Nano-Electro-Mechanical (NEM) Relay Devices and Technology for Ultra-Low Energy Digital Integrated Circuits

Rhesa Nathanael

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2013-45
http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-45.html

May 1, 2013

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 MAY 2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Nano-Electro-Mechanical (NEM) Relay Devices and Technology for Ultra-Low Energy Digital Integrated Circuits	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California at Berkeley,Electrical Engineering and Computer Sciences,Berkeley,CA,94720**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Complementary-Metal-Oxide-Semiconductor (CMOS) technology scaling has brought about an integrated circuits (IC) revolution over the past 40+ years, due to dramatic increases in IC functionality and performance, concomitant with reductions in cost per function. In the last decade, increasing power density has emerged to be the primary barrier to continued rapid advancement in IC technology, fundamentally due to non-zero transistor off-state leakage. While innovations in materials, transistor structures, and circuit/system architecture have enabled the semiconductor industry to continue to push the boundaries, a fundamental lower limit in energy per operation will eventually be reached. A more ideal switching device with zero off-state leakage becomes necessary. This dissertation proposes a solution to the CMOS power crisis via mechanical computing. Specifically, robust electro-mechanical relay technologies are developed for digital circuit application. A 4-Terminal (4T) relay design is firstly developed. Key technology features include tungsten contacts for high endurance; low-thermal-budget p+-poly-Si0.4Ge0.6 structure for post- CMOS process compatibility; Al2O3 as a reliable insulation material; dry release step to mitigate stiction; and folded-flexure design to mitigate the impact of residual stress. Fabricated relays show good conductance (RON < 10 kΩ), abrupt switching behavior (sub-threshold swing below 0.1 mV/dec), and virtually zero leakage (IOFF ~ 10-14 A). Switching delay in the 100 ns range and endurance exceeding 109 on/off cycles is achieved with excellent device yield (> 95%). With relay design and process optimizations, pull-in voltage below 10 V with less than 1 V hysteresis is achieved. Miniaturization reduces the device footprint to 35μm?50μm, ~10% of the first generation device footprint (120μm?150μm). Relays with multiple source/drain electrodes and multiple gate electrodes are proposed for increased circuit functionality and reduced device count. Finally, simple relay-based logic circuits are demonstrated to show pathways to relaybased digital integrated circuits. The complementary inverter is the basis for all digital logic circuits and is investigated in depth. Relay-based logic gates are demonstrated using CMOS-like and relay-specific design approaches. Multi-input/multi-output relays are proposed to enable any complex logic function to be implemented compactly with only two relays.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**136**	

Nano-Electro-Mechanical (NEM) Relay Devices and Technology for Ultra-Low Energy Digital Integrated Circuits

by

Rhesa Nathanael

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

and the Designated Emphasis

in

Nanoscale Science and Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Tsu-Jae King Liu, Chair
Professor Elad Alon
Professor Ronald Gronsky

Fall 2012

Nano-Electro-Mechanical (NEM) Relay Devices and Technology for Ultra-Low Energy Digital Integrated Circuits

by

Rhesa Nathanael

Abstract

Nano-Electro-Mechanical (NEM) Relay Devices and Technology
for Ultra-Low Energy Digital Integrated Circuits

by

Rhesa Nathanael

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

Designated Emphasis in Nanoscale Science and Engineering

University of California, Berkeley

Professor Tsu-Jae King Liu, Chair

Complementary-Metal-Oxide-Semiconductor (CMOS) technology scaling has brought about an integrated circuits (IC) revolution over the past 40+ years, due to dramatic increases in IC functionality and performance, concomitant with reductions in cost per function. In the last decade, increasing power density has emerged to be the primary barrier to continued rapid advancement in IC technology, fundamentally due to non-zero transistor off-state leakage. While innovations in materials, transistor structures, and circuit/system architecture have enabled the semiconductor industry to continue to push the boundaries, a fundamental lower limit in energy per operation will eventually be reached. A more ideal switching device with zero off-state leakage becomes necessary.

This dissertation proposes a solution to the CMOS power crisis via mechanical computing. Specifically, robust electro-mechanical relay technologies are developed for digital circuit application. A 4-Terminal (4T) relay design is firstly developed. Key technology features include tungsten contacts for high endurance; low-thermal-budget p^+-poly-$Si_{0.4}Ge_{0.6}$ structure for post-CMOS process compatibility; Al_2O_3 as a reliable insulation material; dry release step to mitigate stiction; and folded-flexure design to mitigate the impact of residual stress. Fabricated relays show good conductance ($R_{ON} < 10$ kΩ), abrupt switching behavior (sub-threshold swing below 0.1 mV/dec), and virtually zero leakage ($I_{OFF} \sim 10^{-14}$ A). Switching delay in the 100 ns range and endurance exceeding 10^9 on/off cycles is achieved with excellent device yield (> 95%). With relay design and process optimizations, pull-in voltage below 10 V with less than 1 V hysteresis is achieved. Miniaturization reduces the device footprint to 35μm×50μm, ~10% of the first generation device footprint (120μm×150μm). Relays with multiple source/drain electrodes and multiple gate electrodes are proposed for increased circuit functionality and reduced device count.

Finally, simple relay-based logic circuits are demonstrated to show pathways to relay-based digital integrated circuits. The complementary inverter is the basis for all digital logic circuits and is investigated in depth. Relay-based logic gates are demonstrated using CMOS-like and relay-specific design approaches. Multi-input/multi-output relays are proposed to enable any complex logic function to be implemented compactly with only two relays.

To my family,
for their unbounded love...

Table of Contents

Acknowledgement

I would like to begin by thanking The Almighty God, my Lord and Savior; for all accomplishments come from Him and done by His will. It is a bittersweet moment to see one chapter of my life ending with the filing of this dissertation. This has truly been an incredible journey! And I certainly did not get here alone. This triumph does not solely belong to me, but to all those individuals who have played their part along the journey.

First and foremost, my utmost gratitude goes to my research advisor, Professor Tsu-Jae King Liu, for her mentorship, thoughtfulness, kindness, support, and encouragement. In my pursuit to find the ideal switch, I have been blessed to find the ideal advisor who truly goes above and beyond in every way. She is someone her students can always turn to at any time of need. Her depth and breadth of knowledge have greatly contributed to guiding my research directions. In fact, it was her seminar almost 10 years ago that I attended by chance, that kindled my interest in integrated circuits devices. She is one of the brightest and most dedicated person I have ever met, who has been a great inspiration for me professionally and in life. It is a great privilege to be able to call her my teacher and mentor. Thank you!

I would like to thank Professor Elad Alon for advising and sharing his wealth of knowledge on circuit design throughout the collaboration we have in demonstrating relay-based circuits, and for serving in my thesis and qualifying exam committee. He has been ever helpful and a great resource I can always turn to. I also thank the other circuit design collaborators, Professor Dejan Markovic (UCLA) and Professor Vladimir Stojanovic (MIT) for a fruitful collaboration over the years. Without their contributions, the NEM relay project would not have accomplished so much.

I thank Professor Ronald Gronsky for serving in my thesis and qualifying exam committee. I have found his classes thoroughly enjoyable and rewarding as well. I would like to thank Professor Chenming Hu for serving in my qualifying exam committee, reader for my M.S. thesis, and the various advice and help he provided me ever since my undergraduate days. His class first introduced me to the field of semiconductor devices. I would like to express my appreciation to Professor Vivek Subramanian who gave me my first taste of research in his group and provided me with guidance and support to transition to graduate school in the beginning.

I am fortunate to have been able to work closely with many of the past and present members of the magnificent NEMS group. Dr. Hei "Anderson" Kam, Dr. Joanna Lai, Dr. Donovan Lee, and Dr. Vincent Pott have all been great mentors and friends as I was first starting in the relay project. My colleagues from the NEM Relay team, Professor Jaeseok Jeon, Dr. Louis Hutin, I-Ru "Tim" Chen, Yenhao "Philip" Chen, Jack Yaung, and Eung Seok Park, have provided various technical and personal help throughout. I highly value the interactions we had and friendship we all share. Also, Forrest Laskowski has diligently helped relay characterization

efforts during the summer he spent at Berkeley. This has truly been a special team full of wonderful personalities!

I would like to also thank my collaborators whose efforts have made the NEM Relay project successful: Matthew Spencer (UC Berkeley), Abhinav Gupta (UC Berkeley), Dr. Fred Chen (MIT), Hossein Fariborzi (MIT), Chengcheng Wang (UCLA), Kevin Dwan (UCLA), Dr. Wieze Xiong (Sematech), Chanro Park (Sematech), and Rinus Lee (Sematech).

I am sincerely grateful to the fine members of King Group, the wider Device Group, Organic Electronics Group, and EECS department in general past and present: students, staff, postdoctoral and visiting scholars. I treasure the interaction we had as colleagues, as well as friends, as our paths crossed at various points of my graduate school experience. Some of them helped me professionally or personally, through action or advice. Some I have learned a thing or two from. Others are simply a joy to interact with and had made my days brighter. I thank especially Professor Changhwan Shin, Dr. Sung Hwan Kim, Dr. Xin Sun, Dr. Reinaldo Vega, Dr. Min Hee Cho, Nattapol Damrongplasit, Dr. Nuo Xu, Dr. Byron Ho, Dr. Zach Jacobson, Wook Hyun Kwon, Dr. Li-Wen Hung, Dr. Darsen Lu, Dr. Cheuk Chi Lo, Sriramkumar Venugopalan, Dr. Sapan Agarwal, Dr. Peter Matheu, Dr. Anupama Bowonder, Dr. Pratik Patel, Dr. Kanghoon Jeon, Dr. Jemin Park, Professor Woo Young Choi, Dr. Tanvir Morshed, Dr. Kyoungsub Shin, Koichi Fukuda, Taro Osabe, Dr. Si-Woo Lee, Dr. Alvaro Padilla, Dr. Sriram Balasubramanian, Dr. Drew Carlson, Dr. Pankaj Kalra, Dr. Varadarajan Vidya, Dr. Chun-Hsun Lin, Dr. Mohan Dunga, Jodie Zhang, Dr. Morgan Young, Dr. Steve Molesa, Dr. Shong Yin, Steve Volkman, Dr. Qintao Zhang, Dr. Kin-Yip Phoa, Dr. Lakshmi Jagannathan, Dr. Tim Bakhishev, Dr. Alejandro de la Fuente, Dr. Jihoon Park, Dr. Renaldi Winoto, and Dr. Lynn Wang. Thank you all!

The relays fabricated in this work would not have materialized if not for the efforts of the ever helpful staff members of the UC Berkeley Marvell Nanofabrication Laboratory in keeping the lab up and running, maintaining equipments, and providing processing assistance and advice. I thank especially Dr. William Flounders, Jimmy G. M. Chang, Joseph Donnelly, Sia Parsa, Evan Stateler, Jay Morford, Eric Chu, Kim Chan, Robert M. Hamilton, Marilyn Kushner, Laszlo Petho, David Lo, Danny Pestal, Madeleine Leullier, Brian McNeil, Susan Kellogg-Smith, Rosemary Spivey, and Adrienne Ruff.

I am grateful to Ruth Gjerde, Dana Jantz, and Shirley Salanio, the graduate matters assistants, who have been great sources of information and provided me with various assistances in fulfilling graduation requirements. They always welcome students with a smile!

Last but not least, I am tremendously grateful to my family for their unconditional love and support. I am highly indebted to my parents, Ridwan Santoso and Lana Budimanta, who continue to make sacrifices for me, shape and lead me into the individual I am today. Their guidance, support, and endless care have continued to propel me forward through life's numerous challenges. I am exceptionally thankful to my brother, Renard Gamaliel, who has been my greatest companion in whom I can place my utmost trust. He has always been there for

me as together we navigate through the many adventures of life. I am truly blessed to be part of this wonderful family!

The work in this dissertation is supported in part by the Defense Advanced Research Projects Agency (DARPA) / Microsystems Technology Office (MTO) Nano-Electro-Mechanical Systems (NEMS) program, and the Microelectronics Advanced Research Corporation (MARCO) Focus Center Research Program (FCRP): Center for Materials, Structures, and Devices (MSD) and Center for Circuit and System Solutions (C2S2).

Chapter 1
Introduction

1.1 CMOS Technology Scaling

The invention of the integrated circuit (IC) in the summer of 1958 by Jack St. Clair Kilby at Texas Instruments [1], for which he received the Nobel Prize in 2000, was a revolutionary innovation that transformed the world into the one we know today. For more than four decades, the world has seen a dramatic increase in computing power following an exponential trend. The co-founder and Chairman Emeritus of Intel Corporation, Gordon Earle Moore, made an observation in 1965 that the number of transistors on a chip approximately doubles every two years [2], [3]. This trend has continued and famously became known as "Moore's Law." The world's first single-chip microprocessor, the Intel 4004 (introduced in 1971), had 2,300 transistors and operated with a clock frequency of 108 kHz. Today's top-of-the-line microprocessor for consumer products, the Intel Core i7 Extreme Processor (2012 edition), has 2.27 billion transistors spread over six processing cores, each operating with a clock frequency of up to 4GHz [4].

This rapid technology advancement has been enabled by the steady miniaturization of the transistor. Transistor scaling results in improved transistor performance and allows a higher degree of integration, leading to more functionality per chip and reduced cost per function [5]. Moore's Law sets the pace of innovation. The International Technology Roadmap for Semiconductors (ITRS) [6] is a manifestation of Moore's Law that sets the IC technology targets which drive semiconductor device and process research and development. Figure 1.1 shows how the physical gate length of a transistor scales over time, as outlined in the ITRS. Although the industry and consumers have become accustomed to and expect the same rate of advancement each year, it is increasingly difficult to sustain this pace due to various device phenomena that emerge at the nanometer scale and power density constraints. Nonetheless, this long-established scaling trend is still expected to continue for at least another decade according to the ITRS.

In the recent decades, the planar bulk silicon Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET) has been the main building block with which ICs are built. MOSFETs with n-type (n-channel) and p-type (p-channel) source/drain regions can be

utilized together to achieve complementary switching behavior, *i.e.* only one device is turned on at a time when the gate voltage is high (V_{DD}) or low (0V). Specifically, a p-channel MOSFET turns on when gate voltage is low to help "pull-up" the output node to V_{DD}, while an n-channel MOSFET turns on when the gate voltage is high to help "pull-down" the output node to ground (GND). A direct current path from V_{DD} to GND is avoided, limiting static power dissipation when the devices are not switching. As the device density in integrated circuits grew exponentially over time, this complementary metal-oxide-semiconductor (CMOS) transistor technology became the technology of choice since it offered lower static power dissipation than other transistor technologies.

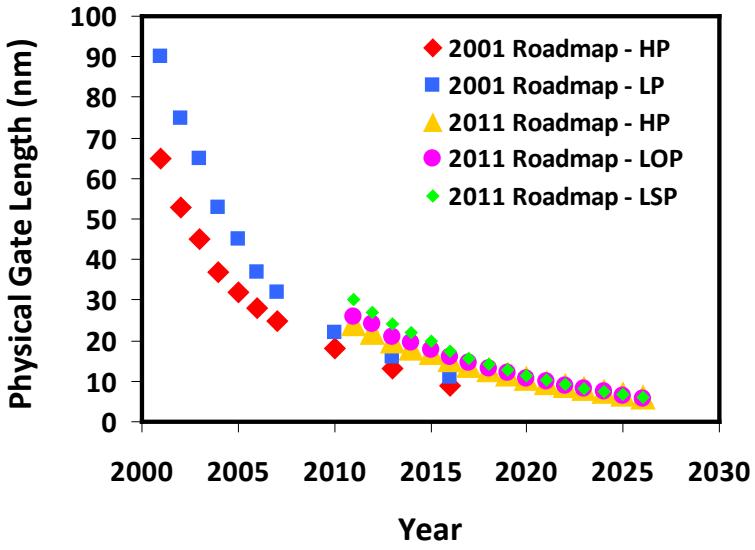


Figure 1.1: Transistor scaling trend as outlined by the ITRS [6]. Physical gate lengths for high performance (HP) and low power (LP) applications are shown.

Originally, simply scaling all device dimensions proportionately had been sufficient to keep improving performance. In the deep-submicron regime, however, an onslaught of new challenges makes scaling increasingly difficult [7]-[9]. These include lithography limitations for printing sub-wavelength features, short-channel effects, increasing static leakage with threshold voltage reduction, increasing variability in transistor performance, *etc*. To continue transistor scaling, the use of novel materials, process techniques, and structures becomes a necessity. While creative solutions have so far been able to alleviate these challenges, a power crisis has emerged due to a more fundamental issue inherent in the operating principle of the MOSFET itself.

While the density of transistors has increased greatly, their operating voltage has not scaled proportionately. Figure 1.2 shows how the CMOS supply voltage (V_{DD}) and threshold voltage (V_{TH}) have scaled with each technology generation. Traditionally, CMOS power consumption was reduced by lowering V_{DD} and appropriately scaling V_{TH} to maintain the same gate overdrive (V_{DD}-V_{TH}) and therefore performance. However, due to

exponentially increasing off-state leakage with V_{TH} reduction, voltage scaling has become limited, especially since the 0.13µm generation of CMOS technology.

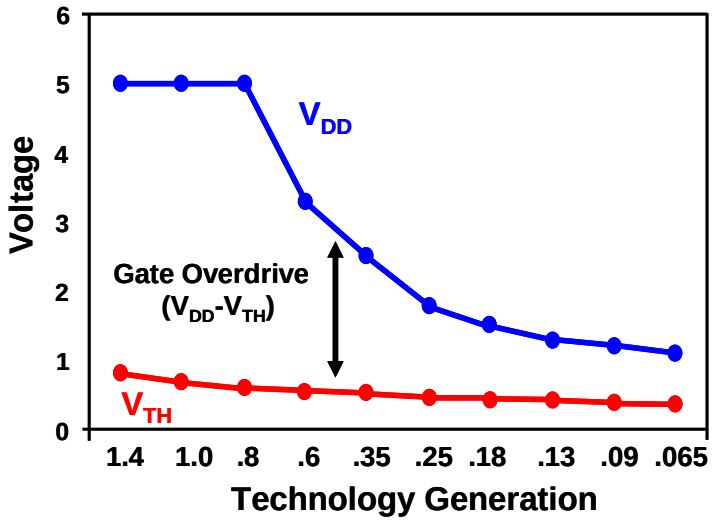


Figure 1.2: Plot of V_{DD} and V_{TH} scaling with technology generation, reproduced from [11]. V_{TH} is no longer being scaled down proportionately with transistor dimensions.

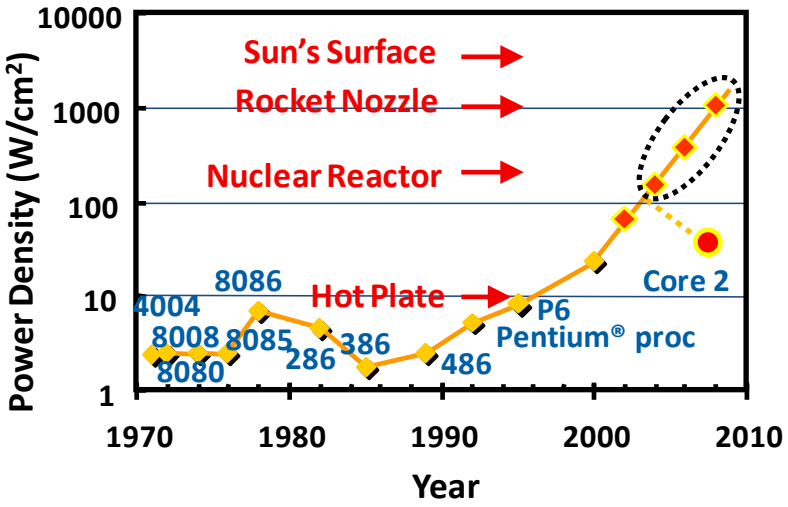


Figure 1.3: Power density for various microprocessor chips, reproduced from [10]. A multi-core approach (Core 2) is adopted to continue to improve system performance within a power density constraint.

Figure 1.3 shows how microprocessor chip power density has increased exponentially over the past decade, a trend that would have led to levels that can no longer be supported by conventional cooling technology. Clearly, power density has become a constraint for chip design. Parallel processing (*i.e.* a multi-core approach) has been employed to prolong scaling trends, but this approach will ultimately be insufficient, due to a fundamental limit in CMOS energy efficiency as described in the next section.

1.2 The CMOS Energy Efficiency Limit

A MOSFET is essentially an electronic switch whose on/off state is controlled by the voltage difference between the gate and the source (V_{GS}). With the channel doped of opposite type as the source/drain regions, a built-in potential barrier is formed at the source-channel junction in the off state. The height of this barrier governs the rate at which mobile charge carriers diffuse from the source into the channel region via thermionic emission and then drift across the channel region to be collected by the drain (*i.e.* the amount of current flowing through the device). By changing V_{GS}, the channel potential and hence the source potential barrier height is modulated, to control the current flowing through the channel.

Consider the typical drain current (I_D) vs. gate voltage (V_G) characteristic of a MOSFET (Figure 1.4(a)). When the magnitude of V_{GS} exceeds the threshold voltage (V_{TH}), the potential barrier at the source side is insignificant so that carriers can easily diffuse into the channel region. In the on state, a conductive path of mobile charge carriers forms in the channel region and current flow is limited by the rate of carrier drift to the drain. The on-current (I_{ON}) of a long-channel MOSFET is given by the following equation:

$$I_{ON} \propto \mu_{eff} C_{ox} \frac{W}{L} (V_{DD} - V_{TH})^2 \qquad (1.2.1)$$

where μ_{eff} is the effective carrier mobility, C_{ox} is the gate oxide capacitance per unit area, and W and L are the transistor gate width and length, respectively.

When I_D is plotted on a logarithmic scale *vs.* V_G (Figure 1.4(b)), it can be clearly seen that the off-state to on-state transition is not abrupt. The energy distribution of carriers within the source region follows Boltzmann statistics. Thus, in the subthreshold ($V_G < V_{TH}$) region of operation, carrier diffusion (and hence I_D) increases exponentially as the potential barrier height is lowered with increasing V_G.

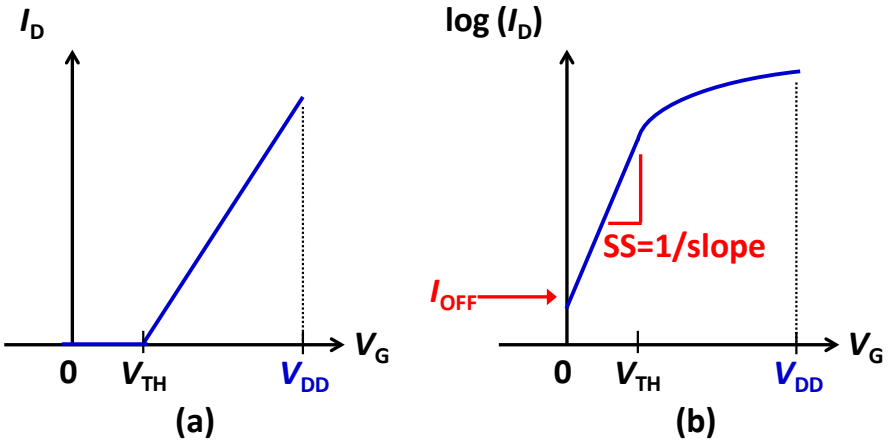


Figure 1.4: Typical plots of drain current (I_D) vs. gate voltage (V_G) for an n-channel MOSFET. I_D is plotted in **(a)** linear scale and **(b)** log scale.

The off-state leakage current (I_{OFF}), *i.e.* the drain current for V_{GS} = 0 V and V_{DS} = V_{DD}, is given by:

$$I_{OFF} \propto 10^{-V_{TH}/SS} \tag{1.2.2}$$

where the subthreshold swing (SS) is defined as the inverse slope of the log(I_D)-V_G curve:

$$SS = \ln(10)\frac{kT}{q}\left(1+\frac{C_{dep}}{C_{ox}}\right). \tag{1.2.3}$$

$\frac{kT}{q}$ is the thermal voltage (26 mV at room temperature), and $\left(1+\frac{C_{dep}}{C_{ox}}\right)$ is a non-ideality factor due to the capacitive voltage divider effect of the semiconductor depletion capacitance (C_{dep}). In the ideal case, C_{ox}>>C_{dep} so that the channel potential closely follows the gate potential and SS reduces to $\ln(10)\frac{kT}{q}$, which is 60 mV/dec at room temperature. Since $\frac{kT}{q}$ is a non-scaling physical constant, an ideal MOSFET switching characteristic is limited to be no steeper than 60 mV/dec at room temperature. In practice, SS is typically closer to 100 mV/dec. The fundamental limit on how abruptly a MOSFET can switch on/off due to the non-scalability of the thermal voltage results in a minimum energy limit for CMOS digital circuits.

The total energy dissipated in a CMOS circuit consists of two components: dynamic energy ($E_{DYNAMIC}$) from charging and discharging capacitors and leakage energy ($E_{LEAKAGE}$) caused by transistor off-state leakage current:

$$E_{TOTAL} = E_{DYNAMIC} + E_{LEAKAGE} \tag{1.2.4}$$

$$E_{DYNAMIC} = \alpha L_D f C {V_{DD}}^2 \tag{1.2.5}$$

$$E_{LEAKAGE} = L_D f I_{OFF} V_{DD} t_{DELAY} \tag{1.2.6}$$

where is α the activity factor, L_D is the logic depth, f is the fanout, and C is the capacitance per stage.

The time delay per operation, t_{DELAY}, is given by:

$$t_{DELAY} = \frac{L_D f C V_{DD}}{2 I_{ON}} \tag{1.2.7}$$

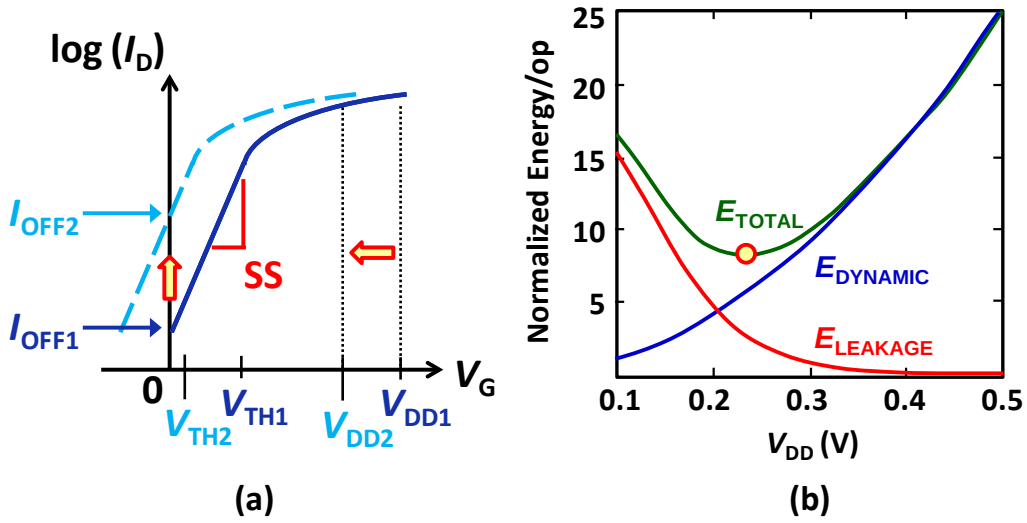


Figure 1.5: CMOS minimum energy limit due to subthreshold leakage. **(a)** log(I_D)-V_G plot showing the effect of lowering V_{DD} and V_{TH} and its **(b)** energy per operation implications.

Dynamic energy can be reduced by lowering V_{DD} following the blue line in Figure 1.5(b). However, by doing so, transistor drive current (I_{ON}) is also reduced (Figure 1.5(a)), which in turn increases delay (t_{DELAY}). In order to maintain the same drive current (*i.e.* circuit performance), the threshold voltage (V_{TH}) must be reduced to maintain the same gate overdrive ($V_{DD} - V_{TH}$). However, a linear reduction in V_{TH} increases the off-state leakage (I_{OFF}) exponentially according to SS, following the red line in Figure 1.5(b). Alternatively, V_{TH} can be held fixed while scaling V_{DD} at the expense of performance (a longer t_{DELAY}), which in turn increases $E_{LEAKAGE}$ as well. Hence, there is a minimum total energy point that balances these two energy components. This is the energy efficiency limit of CMOS that will always exist if the mechanism for on/off switching is thermionic emission over a potential barrier.

To avoid unreasonably high levels of chip power density, parallelism (multi-core processing) has been adopted in recent years. The idea is to operate the circuits more slowly at a lower energy point (Figure 1.6(a)), and to run multiple processor cores in parallel to recoup system level performance (Figure 1.6(b)). Today, multi-core microprocessor chips (with up to 8 cores in a single chip) have become the norm.

Parallelism is only a temporary fix, however. Due to the fundamental energy efficiency limit for CMOS, energy consumption per operation cannot be lowered indefinitely by reducing core performance. When the CMOS circuitry is operating at the minimum energy point, the energy per operation cannot be lowered anymore even if throughput is further reduced (Figure 1.6(c)). To continue to improve system performance in the long run, subthreshold leakage needs to be eliminated. A new switch will be required.

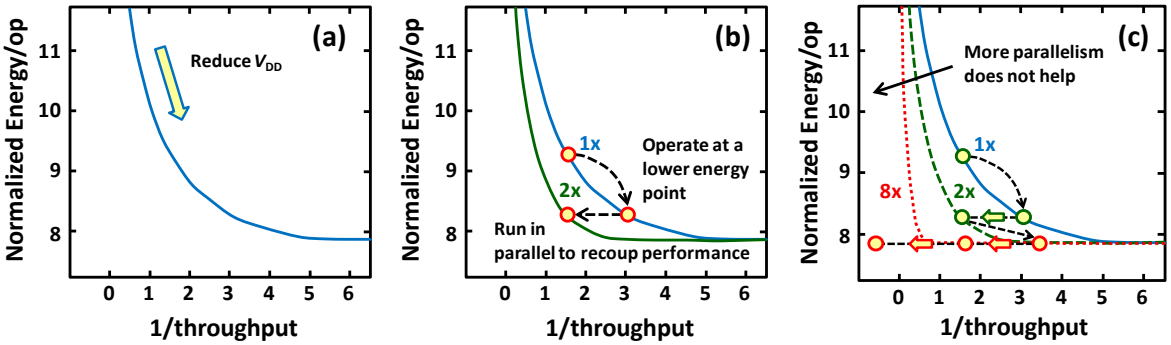


Figure 1.6: Plots of normalized energy per operation *vs.* 1/throughput demonstrating the benefits of parallelism and its limited effectiveness due to CMOS minimum energy point.

1.3 Mechanical Switches

1.3.1 Breaking the Energy Efficiency Limit

A solution to the power crisis that CMOS technology faces today requires a device that has more ideal switching behavior. An ideal switch has very abrupt on/off switching transition and no leakage current in the off-state. Is there such a device?

Behold the mechanical switch! The use of mechanical switches for digital computing is, in fact, not a new concept. They were used long before CMOS technology came into existence. Back then, mechanical switches were relatively large and slow. They were abandoned with the integrated circuits revolution in the mid-1900s, due to functionality and cost benefits of monolithic integration on a single chip. With modern planar processing technology, it is possible today to fabricate miniature relays and integrate them on a single chip, as well. A return to the computing device of old suddenly looks intriguing.

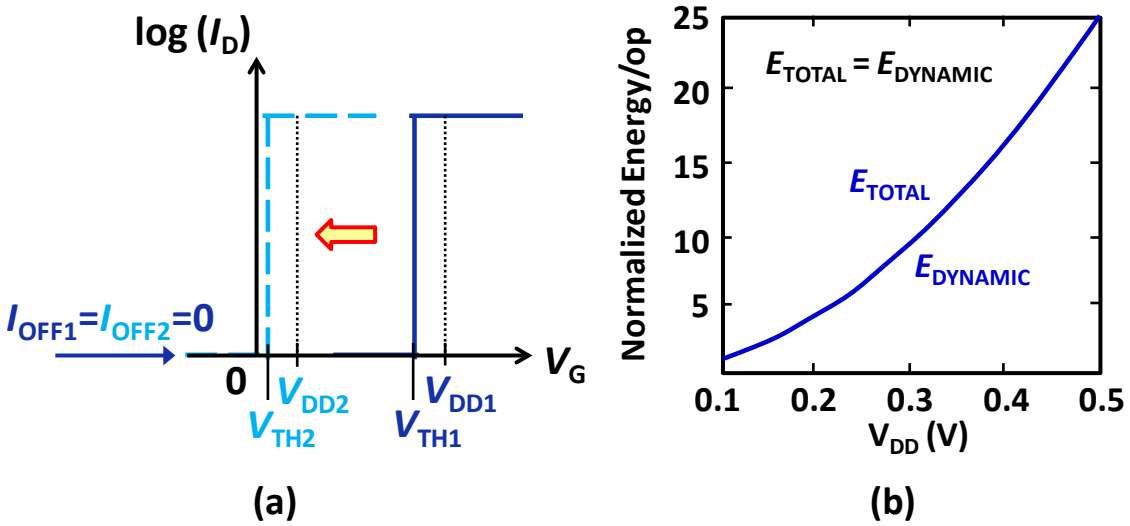


Figure 1.7: Energy efficiency illustration for a mechanical switch. **(a)** I_D-V_G plot for a mechanical switch, showing the effect of lowering V_{DD} and V_{TH} and its **(b)** energy per operation implications.

Figure 1.7(a) shows representative I_D-V_G curves for mechanical switches, which operate based on making and breaking physical contact. There is no modulation of a potential barrier. Thus, the *I-V* characteristic is extremely abrupt (a step-like function). In the off-state, an air gap separates the electrodes, so that off-state leakage is effectively zero ($I_{OFF} = 0$). Therefore, the leakage energy term drops out completely:

$$E_{LEAKAGE} = 0 \tag{1.3.1.1}$$

Thus, the total energy dissipated in a mechanical digital circuit consists only of dynamic energy for charging and discharging capacitances:

$$E_{TOTAL} = E_{DYNAMIC} = \alpha L_D f C {V_{DD}}^2 \tag{1.3.1.2}$$

Without a lower bound imposed by leakage, the energy per operation for a mechanical logic technology could be reduced to be less than the minimum energy per operation for CMOS technology, by simply scaling V_{DD}. This is because, in theory, V_{DD} and V_{TH} can be lowered to be close to 0V for an ideal switch.

1.3.2 Electrostatic Relays

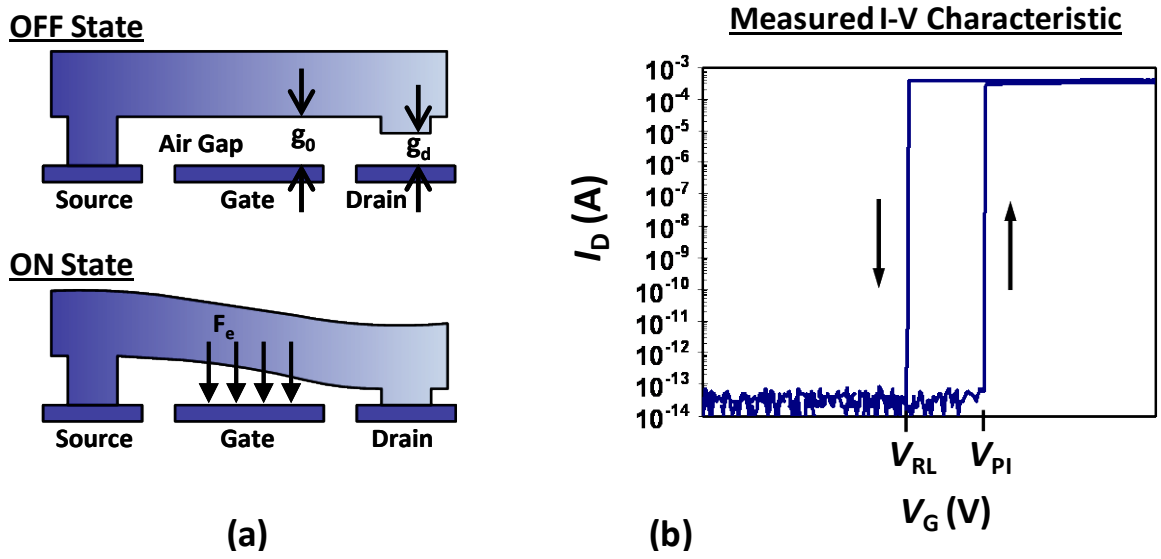


Figure 1.8: **(a)** Schematic illustrations of a generic 3-Terminal relay structure in the off-state and the on-state. **(b)** A typical I_D-V_G characteristic, showing the pull-in (V_{PI}) and release (V_{RL}) voltages.

Micro-electro-mechanical devices utilizing electrostatic actuation are attractive because they are relatively easy to manufacture using conventional planar processing techniques and materials, do not consume substantial active power, and are more scalable compared to other types of actuators such as magnetic, piezoelectric, and thermal. In an electrostatic switch, a movable electrode and a fixed electrode form a capacitor. When a voltage is applied to this capacitor, electrostatic force (due to attraction between oppositely charged electrodes) accelerates the movable electrode toward the fixed electrode. This force is always attractive regardless of the polarity of the applied voltage (*i.e.* it is ambipolar), and its strength depends on the area of the electrodes and the separation between them (the actuation gap, g).

A schematic of a basic 3-Terminal (3T) relay is shown in Figure 1.8(a). In the off-state, the source and drain are physically separated by an air gap, so that no current can flow. When voltage (*i.e.* electrostatic force) is applied between the gate and source, the structure is actuated downward. When the source and drain are in contact, current can flow from drain to source and hence the relay is on. A typical I_D-V_G curve of an electrostatic relay is shown in Figure 1.8(b). A relay turns on when the gate voltage exceeds the pull-in voltage (V_{PI}) and turns off when the gate voltage is lowered below the release voltage (V_{RL}). The switching behavior is abrupt and the off-state leakage is zero.

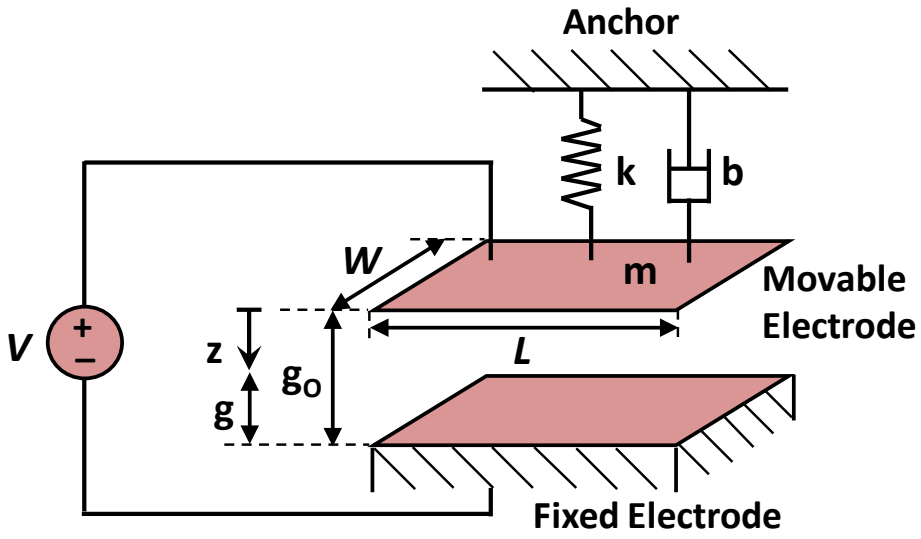


Figure 1.9: Model of an electrostatically actuated beam as a dynamic parallel-plate capacitor and mass-spring-damper system.

An electrostatically actuated beam can be modeled as a parallel plate capacitor on a mass-spring-damper system (Figure 1.9). The motion dynamics of the movable electrode follows Newton's second law of motion:

$$m\ddot{z} + b\dot{z} + kz = F_{elec}(z) \qquad (1.3.2.1)$$

where z is the displacement of the movable electrode, m is the mass of the movable electrode, b is the damping factor, and k is the spring constant. F_{elec} is the electrostatic force when a voltage, V, is applied between the two electrodes, and is a function of the displacement, z. A full solution of the non-linear differential equation needs to be solved numerically.

For the purpose of gaining a fundamental understanding of the operation of an electrostatic actuator, we can model the structure as a simple parallel plate capacitor [12]. The electrostatic force is

$$F_{elec}(z) = \frac{\varepsilon_0 (WL) V^2}{2(g_0 - z)^2} = \frac{\varepsilon_0 (WL) V^2}{2g^2} \quad (1.3.2.2)$$

where g is the actuation gap thickness, g_0 is the as-fabricated actuation gap thickness, W is the width of the actuation area, and L is the length of the actuation area.

The opposing force is the spring restoring force:

$$F_{spring}(z) = kz = k(g_0 - g) \quad (1.3.2.3)$$

At equilibrium, these two forces must balance:

$$\frac{\varepsilon_0 (WL) V^2}{2g^2} = k(g_0 - g) \quad (1.3.2.4)$$

Note that F_{elec} increases superlinearly with displacement, while F_{spring} increases linearly with displacement. Therefore, there exist a point beyond which F_{elec} is always greater than F_{spring}, and the system becomes unstable. By stability analysis of Equation 1.3.2.4, we can show that this critical displacement occurs when the electrode has moved by $\frac{1}{3} g_0$ (*i.e.* $g = \frac{2}{3} g_0$). At this critical point, the gap closes abruptly as the applied voltage is increased. This phenomenon is referred to as "pull-in" and the voltage at which it occurs is the pull-in voltage (V_{PI}), given by:

$$V_{PI} = \sqrt{\frac{8k{g_0}^3}{27\varepsilon_0 (WL)}} \quad (1.3.2.5)$$

Note that the relay design in Figure 1.8(a) employs a dimple at the contacting region to precisely define the apparent contact area. As a result, a smaller air gap exists at the dimple region (g_d) than the actuation region (g_0). Note that if $g_d < \frac{1}{3} g_0$, the pull-in phenomenon will not occur. The thicknesses of these two gaps can be defined independently. Thus having a dimple allows the relay's mode of operation (pull-in or non-pull-in) to be defined by adjusting the ratio of the two gap thicknesses.

For a more general (arbitrary) relay design, the turn-on voltage in pull-in mode can be expressed as:

$$V_{PI} = \sqrt{\frac{8k_{eff}{g_0}^3}{27\varepsilon_0 A_{eff}}} \qquad \text{when } g_d \geq \frac{1}{3} g_0 \quad (1.3.2.6)$$

where k_{eff} is the effective spring constant of the suspension beam(s) and A_{eff} is the effective overlap area between the movable and fixed electrode.

For a relay designed to operate in non-pull-in mode, contact occurs when $g = g_0 - g_d$. The turn-on voltage can be expressed as:

$$V_{PI} = \sqrt{\frac{2k_{eff}g_d(g_0-g_d)^2}{\varepsilon_0 A_{eff}}} \qquad \text{when } g_d < \frac{1}{3}g_0 \tag{1.3.2.7}$$

Recall that the release (turn-off) voltage (V_{RL}) is lower than V_{PI}. This hysteretic switching behavior is due to pull-in mode operation and surface adhesion force. A relay designed to operate in pull-in mode will snap down once the displacement reaches $\frac{1}{3}g_0$. Once pulled-in, the gap is effectively reduced to $g_0 - g_d$, so the electrostatic force is larger at the same voltage (V_{PI}). Thus, turning off the device requires the voltage to be lowered below V_{PI}. The spring restoring force must overcome both the electrostatic force and surface adhesive force that exists when contact is made, in order to turn off the relay. In non-pull-in mode, only surface adhesive force causes the hysteretic switching behavior, so the hysteresis voltage is expected to be smaller.

Once contact is made, the force balance equation is as follows:

$$\frac{\varepsilon_0(WL)V^2}{2g^2} + F_A = k(g_0 - g) \tag{1.3.2.8}$$

where F_A is the surface adhesive force.

At the release operating point, $g = g_0 - g_d$. The release voltage is found to be:

$$V_{RL} = \sqrt{\frac{2(k_{eff}g_d - F_A)(g_0-g_d)^2}{\varepsilon_0 A_{eff}}} \tag{1.3.2.9}$$

It should be noted that the hysteresis voltage sets the lower limit for relay voltage scaling. Because of its abrupt switching behavior, a relay can in principle be made to operate with V_{RL} close to zero, so that V_{DD} can be reduced to be as low as the hysteresis voltage.

1.4 Dissertation Objectives

This dissertation addresses the CMOS power crisis as follows. Chapter 1 discusses the origin of this crisis and why there exists a fundamental limit to the effectiveness of

parallelism. An alternative switching device with more ideal behavior, the electro-mechanical relay, is proposed to solve this challenge.

Chapter 2 describes fabrication process and materials development work to develop a robust micro-relay technology. Requirements for each material layer, as well as selection and evaluation of materials are discussed. Challenges encountered and possible ways to mitigate them are discussed.

Chapter 3 presents a reliable 4-Terminal relay technology developed for digital logic applications. Performance and reliability characterization results of first-generation logic relays are presented. Non-ideal parasitic effects are identified.

Chapter 4 describes improvements to the 4-Terminal relay technology presented in Chapter 3. A more robust and scalable process is presented. Improved relay designs mitigate parasitic electrostatic effects found in the 1st generation relays. Process and design optimizations are performed to achieve the lowest operating voltage with the process technology available. New relay structures are proposed to increase circuit functionality and provide a means for achieving highly compact relay-based circuits.

Chapter 5 presents relay-based circuit demonstrations. A complementary inverter circuit is characterized and studied in depth to gain insight on the optimal body-biasing schemes for relay logic. All 2-input logic functions are demonstrated using only two relays via multi-input/multi-output design.

Chapter 6 summarizes the findings of this dissertation and its contributions to the field of micro/nanoelectronics. Suggestions for future research are offered.

1.5 References

[1] Jack St Clair Kilby Biography by Texas Instruments. [Online]. Available: http://www.ti.com/corp/docs/kilbyctr/jackstclair.shtml

[2] G. E. Moore, "Cramming more components onto integrated circuits," Electronics, vol. 38, 1965, pp. 114-117.

[3] G. E. Moore, "Progress in digital integrated electronics," in IEDM Tech. Dig., 1975, pp. 11-13.

[4] Intel Processors. [Online] Available: http://ark.intel.com/

[5] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, “Design of ion-implanted MOSFET's with very small physical dimensions,” Journal of Solid-State Circuits, vol. 9, pp. 256-268, 1974.

[6] International Technology Roadmap for Semiconductors. [Online] Available: http://public.itrs.net.
[7] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S.-H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H.-J. C. Wann, S. J. Wind, and H.-S. Wong, "CMOS scaling into the nanometer regime," Proceedings of the IEEE, vol. 85, no. 4, pp. 486-504, 1997.
[8] T. Skotnicki, J. A. Hutchby, T.-J. King, H.-S. P. Wong, and F. Boeuf, "The end of CMOS scaling: toward the introduction of new materials and structural changes to improve MOSFET performance," IEEE Circuits and Devices Magazine, vol. 21, no. 1, pp. 16-26, 2005.
[9] T.-C. Chen, "Overcoming research challenges for CMOS scaling: Industry directions," in Proc. 8th International Conference on Solid-State and Integrated Circuit Technology, (ICSICT'06), pp. 4-7, 2006.
[10] S. Borkar, Intel Corp.
[11] P. Packan, "Device and Circuit Interactions," in IEEE International Electron Device Meeting (IEDM '07) Short Course: Performance Boosters for Advanced CMOS Devices, Dec. 2007.
[12] S. D. Senturia, Microsystem Design. Boston, MA: Kluwer Academic, 2001.

Chapter 2

Relay Process Development

2.1 Introduction

Early computing machines utilized mechanical mechanisms (involving actuation and physical contact). The electromechanical relay was first built by J. Henry in 1835 using an induction coil for actuation, and achieved superior performance than purely mechanical switches. Through the 20th century, computing machines were built using electromechanical devices. However, they were huge in size, slow, and very expensive to build. The era of electronic computing began with vacuum tubes, and took off in the 1950s after the inventions of transistors and integrated circuits. Today, digital computing devices utilize solid-state electronic switches (*i.e.* CMOS transistors). The ability to miniaturize and integrate billions of them on a single chip of silicon the size of a coin has enabled dramatic increases in functionality and performance, as well as reductions in cost [1].

As transistor scaling continues, CMOS technology approaches a fundamental energy efficiency limit. As described in Chapter 1, a shift back to the computing device of old, the relay, potentially can help to overcome this limit. By leveraging advancements over the past 40+ years in planar processing and micro-machining technology, electromechanical relays can be miniaturized to overcome their historical disadvantages.

Micro-electro-mechanical systems (MEMS) comprise miniature mechanical or electromechanical structures fabricated using micro-machining techniques derived from planar processing techniques. As the field grew, specialized processes were developed to meet the unique needs of MEMS structures with materials and structural aspect ratios not typical for CMOS devices. (Interestingly, integrated circuits may borrow processing techniques from MEMS as advanced transistor structures are adopted. For example, deep reactive ion etching (DRIE) techniques originally developed for MEMS to produce highly anisotropic, steep sidewalls, can be used to fabricate "three-dimensional" multi-gate transistors which require high-aspect-ratio silicon fins.)

With the advancement of micromachining technology, MEM relays can be developed to combine the benefits of scaling and the ideal current-*vs*.-voltage characteristic of mechanical switches. Various 3-Terminal (3T) micro-relay designs for logic

applications have been described in the literature. Among them are relays employing cantilever beam designs [2], [3], [4] and clamped-clamped beam designs [5], [6]. The actuated structure can be designed to move vertically (out-of-plane) [2], [4], [5] or laterally (in-plane) [3], [6]. Various contact materials have been explored, including W [5], TiN [2]-[4], and Ru [6]. So far, the lowest operating voltage reported is ~5 V, with a switching delay in the range of hundreds of ns [5], [6]. Poor reliability due to stiction and welding at the contacts remains a key issue [2]-[4], [6]. Coating one of the contact electrodes with a thin SiO_2 layer can improve endurance to over 400 switching cycles [2]. Another method proposed in [4] to attain better reliability is to encapsulate the relay in an insulating liquid medium, such as transformer oil. Endurance was improved from 10 to over 50 cycles using this approach, but at the cost of degraded switching speed. A laterally actuated switch design with Ru contacts achieved 2 million switching cycles [6], but this is still insufficient for logic switch applications.

Monolithic integration of MEMS with CMOS on a single chip is also attractive to enhance system functionality and performance by leveraging the advantages of each technology where appropriate [7]. Integration with CMOS can be achieved with either a "MEMS-first" (MEMS fabrication done before CMOS process steps) or a "MEMS-last" (MEMS fabrication done after CMOS process steps) approach. The "MEMS-first" approach is less desirable because it requires a customized CMOS process. A "MEMS-last" process allows a standard foundry CMOS process to be used and also allows for vertical stacking of MEMS on top of electronics, which saves area and reduces parasitic interconnect resistance and capacitance for better performance; however, it imposes a thermal budget constraint for MEMS fabrication [8].

In this chapter, early process development work to fabricate electrostatically actuated micro-relays for digital logic application is presented. These efforts laid the groundwork for the robust, CMOS-compatible 4-Terminal (4T) relay technology presented in Chapters 3 and 4.

2.2 Relay Structures

A logic relay design is MOSFET-inspired. Terminology such as "source," "drain," "body," and "gate" are used to denote terminals that are analogous to their MOSFET counterparts. The 4-Terminal (4T) relay structure is most useful for circuit applications and is the ultimate goal of this process development effort. Simpler structures have been used along the way to solve process challenges. Trading off simplicity and functionality, each structure is useful for different purposes. Figure 2.1 shows layout and cross-sectional views of the different structures. Their circuit symbols are shown in Figure 2.2.

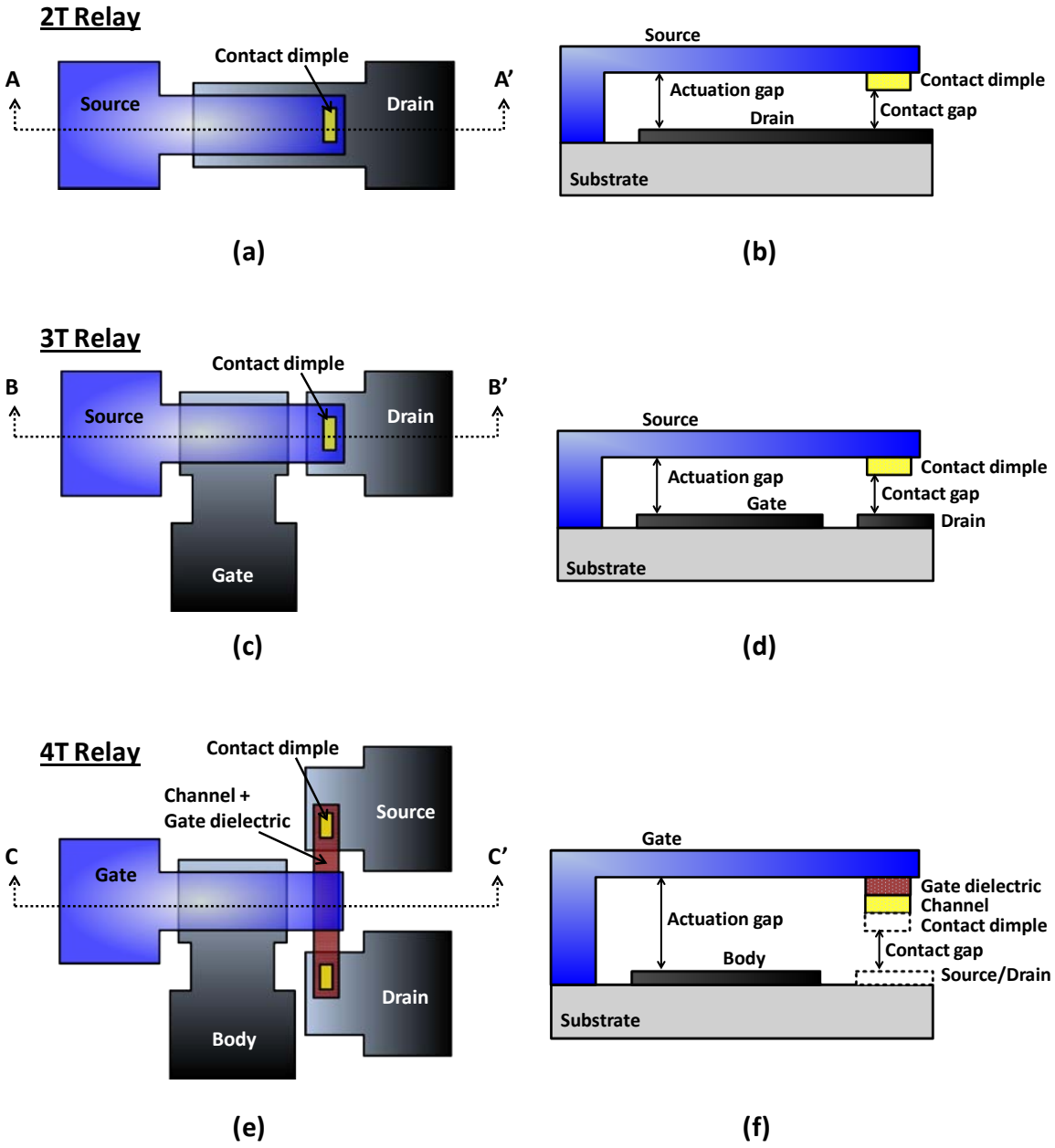


Figure 2.1: Various relay structures used for process development. 2-Terminal relay (a) top view and (b) cross sectional view along A-A'. 3-Terminal relay (c) top view and (d) cross sectional view along B-B'. 4-Terminal relay (e) top view and (f) cross sectional view along C-C'. The contact dimple and source/drain electrodes of the 4-Terminal relay design are not along the cross section but are shown with dotted lines to indicate their heights.

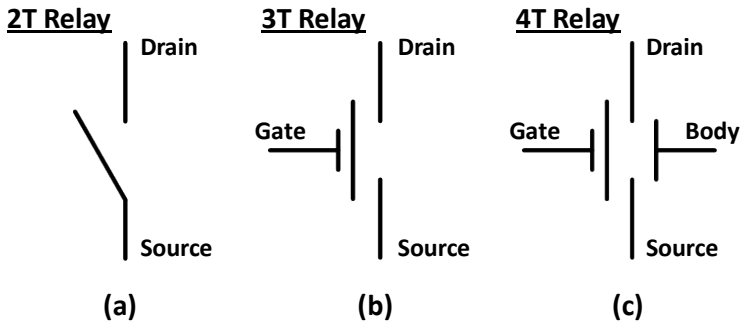


Figure 2.2: Circuit symbols for (a) 2-Terminal, (b) 3-Terminal, and (c) 4-Terminal relays.

2.2.1 2-Terminal Relay

The most basic relay structure is a simple 2-Terminal (2T) switch illustrated in Figures 2.1(a) and 2.1(b). It simply consists of a source terminal and a drain terminal. A voltage difference between the two terminals actuates the relay and current flows between them when contact is made. Since the actuation voltage and the drain voltage are the same, reliability becomes a major issue. Actuation voltages for mechanical switches tend to be high (~10 V or higher), especially when the technology has not reached maturity. Relays become prone to welding-induced stiction at the contact due to Joule heating from excessively high current flow. Additionally, a 2T switch is not very useful for implementation of digital logic circuits because it does not have separate input (control) and output electrodes.

Despite the shortcomings mentioned above, the 2T switch is useful for process development due to its simplicity. It requires only a single lithography mask to pattern the structural layer to form the movable terminal; the substrate serves as the other terminal. This is sufficient to allow the mechanical properties of the structural layer to be quickly characterized. For example, V_{PI} of a cantilever array can be obtained to extract Young's modulus of the material [9], which relates to structural stiffness. The effects of residual stress and strain gradient on V_{PI} can also be studied: an interferometer measurement can be used to measure the amount of structural deflection resulting from a non-zero strain gradient.

2.2.2 3-Terminal Relay

The 3-Terminal (3T) relay design (Figures 2.1(c) and 2.1(d)) solves the main disadvantage of the 2T relay design by providing separate input (gate) and output (drain)

electrodes. The actuated structure is electrically connected to the source electrode, and its position is determined by the voltage difference between the gate and the source terminals (V_{GS}), which can be high without affecting drain current. Drain current flow (I_{DS}) is determined by the drain-to-source voltage (V_{DS}), which can be small to prevent welding. From a processing standpoint, this allows the electrode and structural materials to be optimized separately. The electrode material can be optimized for good contact reliability while maintaining sufficiently low on-resistance (R_{ON}) for the target application. The structural layer can be optimized for low V_{PI}, for low-power operation.

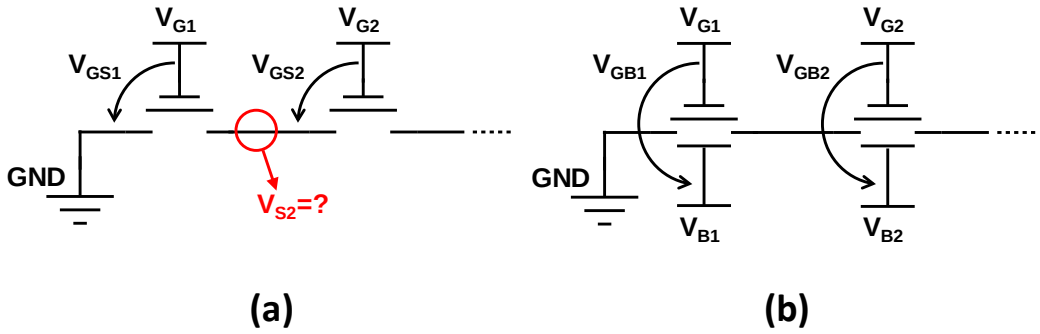


Figure 2.3: Relays connected in series. (a) 3-Terminal relays. (b) 4-Terminal relays always have a fixed gate switching voltage.

The main disadvantage of the 3T relay design is seen from a circuit perspective. In an integrated circuit, multiple relays may be connected in series. (Examples include a conventional NAND or NOR gate.) For 3T relays connected in series, shown in Figure 2.3(a), the gate switching voltage can vary for the relay on the right, since its source voltage is not fixed. This can lead to unreliable circuit behavior.

Nonetheless, a 3T design suffices for the purpose of studying switching behavior for process technology development. Unlike the 2T relay, the 3T relay can turn off reliably since the contact material and drain bias can be optimized for reliable operation. The 3T design is therefore appropriate to study contact properties, such as surface adhesive forces and how the contact resistance evolves with on/off cycling. This is especially useful for choosing and optimizing contact materials, or studying thin-film electrode coatings.

2.2.3 4-Terminal Relay

The 4-Terminal (4T) relay design addresses the shortcoming of the 3T relay design with the addition of the body terminal (Figures 2.1(e) and 2.1(f)) and gate insulating layer.

In this design, the actuated structure is the gate, whose position is controlled by the voltage applied between the gate and the body (V_{GB}), as opposed to the gate and source (V_{GS}) in the 3T design. Thus, the gate switching voltage can be fixed and independent of the source voltage, as shown in Figure 2.3(b). A metallic channel attached to the underside of the gate serves as a bridge connecting the source and drain when the relay is in the on-state. A layer of gate dielectric insulates the gate and the channel so that current flows only from the drain to the source during the on state, not to the gate. Apart from fixing the gate switching voltages, the 4T relay allows the gate switching voltage to be tuned by applying a body bias voltage, which can be advantageous for circuit design. An in-depth discussion of the properties of the 4T relay design is presented in Chapter 3.

2.3 Relay Process Flow

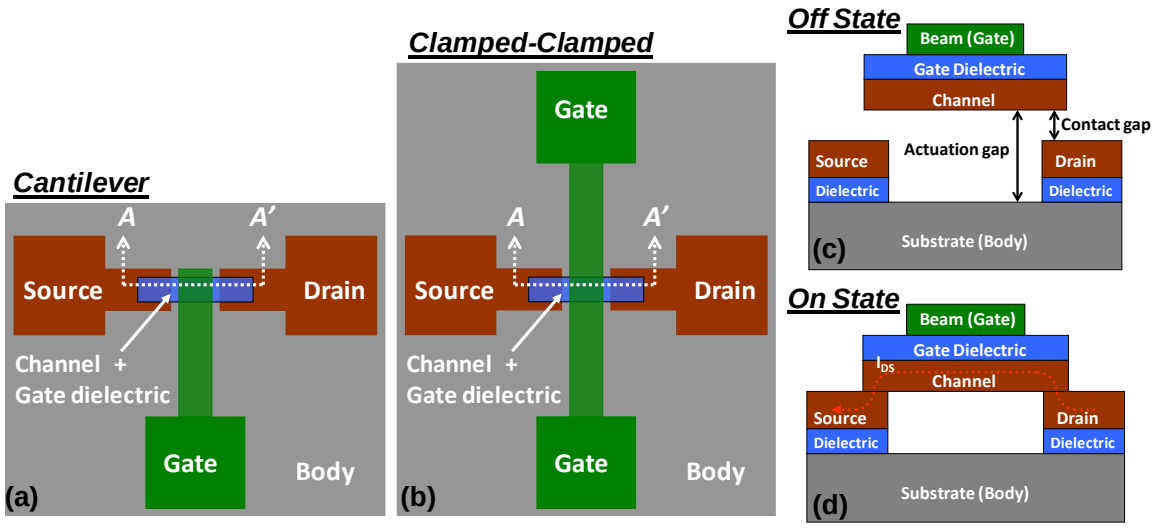


Figure 2.4: Four-terminal relay structure and operation. **(a)** Cantilever beam and **(b)** clamped-clamped beam design layout. **(c)** Cross-section along the channel (A-A') in the off-state and **(d)** on-state.

In this work, the ultimate goal is to develop a robust and reproducible 4T relay technology that can be used to fabricate complex digital circuits. Arrays of cantilever beam and clamped-clamped beam relay designs (Figure 2.4) of varying lengths and widths are used to develop the process. Beam widths range from 1 µm to 3 µm. Beam lengths range from 5 µm to 60 µm and from 11 µm to 120 µm for cantilever and clamped-clamped designs, respectively.

A process flow used to develop a 4T relay technology is illustrated in Figure 2.5. Starting with an n-type silicon substrate, a layer of phosphosilicate glass is deposited at 450°C by LPCVD, then annealed at 1050°C to drive in the dopants. The PSG is then removed in a 5:1 buffered hydrofluoric acid (BHF) solution. The substrate is now heavily n-type doped. A layer of low temperature oxide (LTO) is deposited at 400°C via LPCVD as both an insulating layer for the electrodes and 1st sacrificial layer. Amorphous silicon (A-Si) buffer layer is deposited at 500°C to promote adhesion of the tungsten electrode layer. The tungsten layer is deposited via sputtering and patterned to form source and drain electrodes. A layer of 2nd sacrificial LTO is deposited at 400°C by LPCVD. Next, tungsten followed by an A-Si adhesion layer are deposited and patterned to form the channel. A thin (<10 nm) layer of LTO is then deposited as gate dielectric. (Later Al_2O_3 deposited by atomic layer deposition (ALD) was used as the gate dielectric to improve reliability and solve gate leakage current issues.) The n-type polycrystalline silicon (poly-Si) structural layer is deposited at 610°C by LPCVD. P-type polycrystalline silicon germanium (poly-SiGe) deposited by LPCVD at 410°C is a low thermal budget alternative for the structural layer. An LTO hard mask is deposited at 400°C before patterning the poly-Si or poly-SiGe to form the structure. Finally, the device is released in HF vapor to selectively remove the SiO_2 sacrificial layers.

Note that in this process, the heavily doped n-type substrate serves as the body terminal. Amorphous silicon buffer layers are required because of poor adhesion between tungsten and SiO_2. A contact dimple design is not used, in order to minimize the number of lithography steps. The HF vapor release process is a very carefully timed etch because SiO_2 comprises the gate dielectric as well as the substrate dielectric which supports the source/drain electrodes and the structural anchors. We relied on etch-rate variation with gap thickness and also used large anchor and electrode areas to ensure that the SiO_2 is not completely removed under these areas during the release etch. The next sections describe the selection and development of these materials and the challenges encountered in developing this process.

This process yields functional devices that demonstrate very steep switching behavior, as expected. However, device yield is poor (~50-60%). The cantilever beams are prone to stiction and the clamped-clamped beams are buckled and generally show one-time switching behavior. For the functional devices, endurance was found to be rather poor (<10 cycles). Once they switch, the relays either get permanently stuck down or structurally failed. Nevertheless, valuable lessons were learned to guide efforts toward the robust, reliable, and repeatable process flow discussed in Chapters 3 and 4.

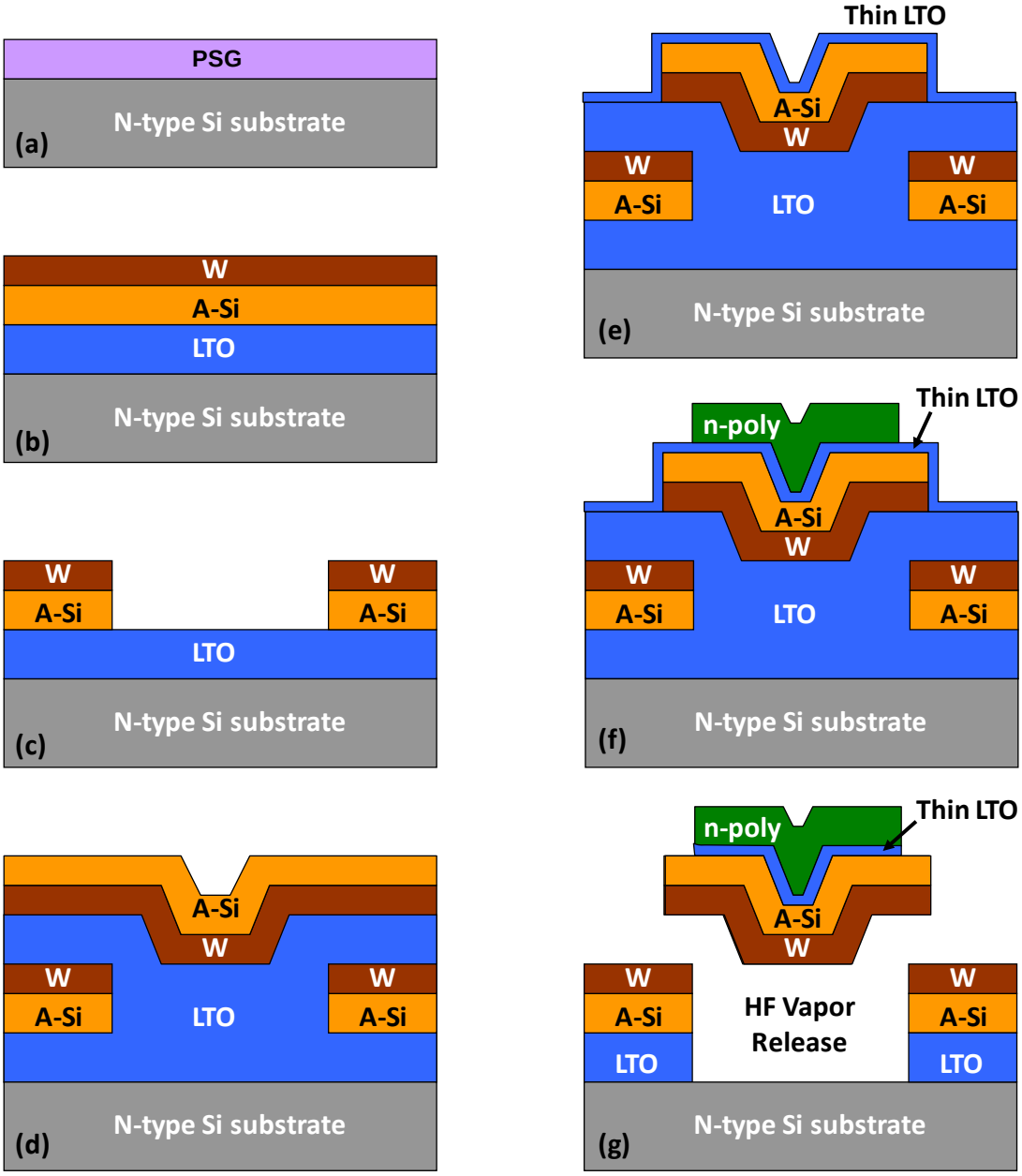


Figure 2.5: Four terminal relay process flow, illustrated with schematic cross-sections corresponding to A-A' in Figure 2.4. **(a)** N-type silicon substrate is heavily doped by PSG predeposition and drive-in anneal. Phosphosilicate glass (PSG) is then removed. **(b)** 1st sacrificial SiO_2 layer is deposited, followed by amorphous silicon (A-Si) buffer layer and tungsten (W) electrode layer. **(c)** Source/drain electrodes are patterned. **(d)** 2nd sacrificial SiO_2 layer deposited. W channel layer and A-Si buffer later deposited. **(e)** Channel is patterned followed by thin SiO_2 deposition for gate oxide. **(f)** N-doped polycrystalline silicon or silicon-germanium is deposited and patterned to form the structure. **(g)** Sacrificial layers are removed by HF vapor release.

2.4 Materials Selection and Development

Developing an integrated device fabrication process involves many considerations. There are four critical elements of the 4T relay structure, namely the sacrificial, contact electrode, insulating dielectric, and structural materials. Each of these has its individual requirements, and also must be process-compatible with the other materials. In this section, selection of these materials are discussed, with considerations related to process integration as well as the application (digital ICs), and availability of materials and processes in the UC Berkeley Marvell Nanofabrication Laboratory.

2.4.1 Sacrificial

In selecting a set of relay materials, the sacrificial material is the first to be considered because it determines the etch chemistry that will be used to release the relay at the very end of the process. This means that every material in the relay structure will be exposed to that particular etch chemistry and must be able to withstand it very well.

The sacrificial layer has the following key requirements:

(1) Dry release process
(2) Uniform and highly controllable deposition
(3) Low deposition temperature (<425°C)

MEMS structures released in a typical wet chemical etch process are prone to stiction due to capillary forces. Surface tension from the liquid etchant can cause adhesion forces stronger than the spring restoring force of the beam, causing the structure to get stuck down during release. More hydrophilic contact surfaces are more susceptible to this type of stiction. Solutions such as critical-point drying to eliminate surface tension have been proposed [10] but add to process complexity and do not solve the problem completely. A dry release process is desirable because it eliminates the presence of fluid during the etch process and thereby mitigates stiction.

Uniformity and controllability of deposition is an important consideration for choosing a sacrificial material. The sacrificial layers determine the actuation and contact gap thicknesses, which in turn determine the switching voltages of the relay. It is critical that the gap thickness can be well-controlled and repeatable during processing, and is uniform across the wafer to ensure uniform operating voltages. The sacrificial layers also must be continuous and without pinholes that may result in shorts between electrodes. In the current process, sacrificial layer thicknesses are ~100 nm. As relay dimensions are scaled down for improved device density and performance, the ability to controllably deposit thin sacrificial films in the nanometer range will be critical.

Ultimately, a CMOS-compatible process is desired. While the ultimate goal is to demonstrate purely relay-based complex logic circuits, integration with CMOS devices would allow the advantages of each technology to be leveraged to realize hybrid systems with enhanced performance and functionality. A "MEMS-last" process is most attractive because it allows CMOS transistors to be fabricated using a conventional process before the relays are fabricated. However, this approach constrains the relay thermal process budget. The thermal budget limit is 425°C for 6 hours or 450°C for 1 hour, for a foundry 0.25 μm CMOS technology [8].

Considering these criteria and materials availability, SiO_2 seems to be a good candidate. It can be selectively removed in HF vapor by suspending the chip on a temperature-controlled electrostatic chuck over a 49% HF bath. SiO_2 can be deposited via low-pressure chemical vapor deposition (LPCVD) with good uniformity and conformality. A low temperature oxide (LTO) process is developed to deposit SiO_2 at 400°C and 300 mTorr process pressure with gas flow rates as follows: SiH_4 = 90 sccm, O_2 = 135 sccm. The SiO_2 deposition rate is ~10 nm/min with good thickness control in the 10s of nm range.

2.4.2 Contact Electrode

The contact electrode material must meet the following requirements:

(1) Sufficiently low contact resistance
(2) Resistant to wear and plastic deformation (high hardness)
(3) Resistant to HF vapor
(4) Uniform deposition process
(5) Low deposition temperature (<425°C)

The electrode material is important because it determines the on-resistance of the relay (R_{ON}) as well as contact reliability. In a 4T relay, R_{ON} consists of several components:

$$R_{ON} = R_{SOURCE} + R_{DRAIN} + R_{CHANNEL} + 2R_{CONTACT} \quad (2.4.2.1)$$

where R_{SOURCE} is the resistance of the source electrode, R_{DRAIN} is the resistance of the drain electrode, $R_{CHANNEL}$ is the resistance of the channel electrode, and $R_{CONTACT}$ is the contact resistance between the channel and source or drain. Out of these components, the contact resistance is dominant, so that $R_{ON} \cong 2R_{CONTACT}$. Therefore, contact properties should be studied carefully.

A resistance model for an electro-mechanical contact has been developed by R. Holm [11]:

$$R_{CONTACT} = \frac{4\rho\lambda}{3A_r} \quad (2.4.2.2)$$

where ρ is the resistivity of the material, λ is the electron mean free path in the material, and A_r is the effective contact area.

The effective contact area is given by:

$$A_r \approx \frac{F_{elec}}{\xi H} \quad (2.4.2.3)$$

where F_{elec} is the electrostatic force that makes the contact (which gives a measure of the loading force), ξ is the deformation coefficient, and H is the hardness of the material.

The contacting surface is always rough [12], so that physical contact is only made at local asperities [13]. The real contact area is only a fraction of the apparent contact area and is a function of the applied load and the material hardness. The applied load should be kept low enough to keep the material in its elastic domain and avoid plastic deformation.

As the field of MEMS emerged 20+ years ago, relays were among the first devices to be investigated. These were built mainly for RF switching applications which require very low on-resistance to minimize insertion loss. As a result, RF switches are typically built with soft metals such as gold (H = 0.2-0.7 GPa), which can achieve very low contact resistance (<1 Ω) [14]. Due to Joule heating from high current flow, welding-induced failure is a major concern for this application. In contrast, low on-resistance is not a requirement for digital logic applications: R_{ON} can be as high as 10 kΩ while ensuring that the electrical charging delay will be much smaller than the mechanical switching delay (~100 ns), for typical load capacitances (10-100 fF) [15]. Logic relays should be designed instead for high reliability to ensure an operating lifetime of 10 years. Hard, refractory metals are therefore more suitable candidates for the contact electrode material.

Tungsten is a good candidate for the contact electrode. It has one of the highest hardnesses of all metals. It has Mohs hardness of ~7.5 and Vickers hardness ~3.43 GPa. Of all pure metals in the periodic table, tungsten has the highest melting point, 3422°C, for good resistance to welding-induced failure due to Joule heating at the contact. Also from a process integration point of view, tungsten is an excellent candidate. First, tungsten is highly resistant to HF, BHF, and HF vapor [16], which is also confirmed in this work. Second, metal thin films in general can be deposited using Physical Vapor Deposition (PVD), such as sputtering. In the UC Berkeley Marvell Nanofabrication Laboratory, tungsten can be deposited with good thickness control in the CPA 9900 DC magnetron sputtering system. Ions from a high-energy plasma are accelerated towards the metal target to bombard and dislodge metal atoms from the target. A magnetron is used to help control the trajectories of these dislodged atoms, which then fall onto the wafer surface and form a film. Deposition is fast with good uniformity and reasonable conformality. Sputtering

easily meets the thermal budget constraint, since the wafer is kept at room temperature. In the CPA system, wafers are placed on a track that transports them across the chamber. Deposition occurs when the wafers pass below the sputtering target. With a faster track speed, the wafers spend less time under the target so that the deposited film is thinner. Prior to the deposition, the chamber is pumped down for a couple of hours to reach base pressure ~1×10^{-7} Torr, so that deposition is carried out under high vacuum. An inert process gas (Ar) flows to the chamber with process pressure = 9 mTorr. Figure 2.6 shows tungsten deposition thickness and sheet resistance for various combinations of power and track speed.

Power	Track Speed	Thickness	Sheet Resistance
1kW	40cm/min	50nm	5Ω/sq
1kW	60cm/min	35nm	7Ω/sq
1kW	80cm/min	30nm	10Ω/sq
1.5kW	40cm/min	55nm	3Ω/sq
1.5kW	60cm/min	50nm	5Ω/sq
1.5kW	80cm/min	40nm	6Ω/sq
2kW	40cm/min	85nm	2Ω/sq
2kW	60cm/min	55nm	3Ω/sq
2kW	80cm/min	50nm	5Ω/sq

Figure 2.6: Tungsten thin-film deposition by DC magnetron sputtering: thickness and sheet resistance for different power and track speed settings. Base pressure is ~1×10^{-7} Torr, and process pressure is 9 mTorr.

2.4.3 Insulating Dielectric

In a 4T relay, a dielectric is needed in two places: 1) insulating layer between the electrodes and the silicon substrate (substrate dielectric), and 2) insulating layer between the channel and the structural electrode (gate dielectric).

The dielectric material has the following requirements:

(1) High electrical breakdown voltage
(2) Low leakage current
(3) Low stress
(4) Resistant to HF vapor
(5) Uniform and controlled deposition process
(6) Low deposition temperature (<425°C)

In relays, the sole purpose of the dielectric layers is electrical insulation. It is important, therefore, to choose a material with high electrical breakdown voltage to be able to withstand the higher operating voltage of MEMS devices. To maintain low static power dissipation and good reliability, gate leakage current should be kept to a minimum. Unlike a CMOS transistor, a relay does not depend on good capacitive coupling between the gate and the channel to operate with steep turn-on/off characteristics, so there is no need for very low equivalent gate-oxide thickness. In fact, a low-permittivity dielectric material is more desirable to reduce capacitive loading. Mechanical stress is a consideration because it can affect the curvature (and hence the actuation and contact gaps) of the released structure. Finally, process integration and thermal budget requirements are considerations.

In the initial stages of process development, the choice of materials is limited. Conventional dielectric materials used in a CMOS process, such as silicon dioxide (SiO_2) and silicon nitride (Si_3N_4), are readily available. Si_3N_4 is uniformly deposited by LPCVD at 800°C, which does not meet the low thermal budget requirement. SiO_2 can be deposited via LPCVD at temperatures between 400°C-450°C. However, it will be etched away during the release step together with the sacrificial SiO_2 layers. It is known that the SiO_2 lateral etch rate in 5:1 BHF and 10:1 HF becomes thickness dependent for gaps < 50 nm [17], due to slower flow rates of etchants and etch by-products. In 5:1 BHF, the lateral etch rate for SiO_2 in a 10 nm gap is about half that in a 50 nm gap. Assuming similar behavior will be observed in HF vapor, a layer of SiO_2 much thinner than the sacrificial layer could potentially survive the release step, if timed carefully.

A low-temperature oxide (LTO) LPCVD process is developed to deposit thin SiO_2 with good control. Deposition occurs at 450°C and 300 mTorr process pressure with gas flow rates as follows: SiH_4 = 1 sccm, O_2 = 9 sccm. The deposition rate is ~1 nm/min with good control in the nm range. Thin (<10 nm) LTO gate dielectric layer is integrated into the process flow and verified to survive the release step after all sacrificial layers are etched away (test structures that mimic the channel are still attached to the structure, indicating that the intermediary thin gate oxide is still present). This is not a very robust solution since the release step must be characterized and timed with precision. Since the sacrificial oxides need to be significantly thicker than the gate oxide to make use of etch rate difference, this process is also not scalable.

The arrival of the PICOSUN SUNALE R150 Atomic Layer Deposition (ALD) tool provided the capability to deposit aluminum oxide (Al_2O_3), a more robust dielectric material. The ALD process involves a surface chemistry of two precursors that forms thin film in a self-limiting manner. The two precursors are pulsed alternately in sequence. With each pulse, surface reaction occurs forming a continuous monolayer of thin film. Unused precursor and reaction by-products are purged in between the pulses. Atomic level precision can be achieved simply by setting the number of precursor pulse cycles. ALD Al_2O_3 can be deposited at 300°C with growth rate of ~0.1 nm/cycle, which easily satisfies the thermal budget requirement.

Al_2O_3 has favorable electrical, mechanical, and chemical properties. A report of ALD Al_2O_3 films formed using the same system (PICOSUN SUNALE R150) shows that a 300 nm film deposited at 300°C has intrinsic tensile stress of ~200 MPa [19], which is within the acceptable range. A film deposited at 250°C shows leakage current density of 1 $\mu A/cm^2$ at 2 MV/cm, with a breakdown at 7.5 MV/cm [19]. In this work, measurements of films deposited at 300°C show leakage current density ~10 nA/cm^2 at 2 MV/cm, while breakdown occurs at electric fields >15 MV/cm. In a device, breakdown is expected to occur at a lower voltage than for a blanket film, due to concentrated electric field at corners formed as a result of topography. Finally, despite being attacked in HF liquid [16], Al_2O_3 is known to be highly resistant to HF vapor [20], which is verified in this work. Al_2O_3 is therefore an excellent relay dielectric material.

2.4.4 Structural

The choice of structural material is an important one since it determines relay operating voltage and mechanical reliability. The important considerations for structural material are the following:

(1) Low residual stress and strain gradient
(2) Robust against fracture and fatigue
(3) Resistant to HF vapor
(4) Uniform and controlled deposition
(5) Low deposition temperature (<425°C)

Polycrystalline silicon (poly-Si) is the conventional structural material used in surface-micromachined MEMS devices [21]. Its mechanical and stress properties have been well characterized for many years [22], [23]. Low residual stress (<25 MPa) with negligible strain gradient can be achieved. However, poly-Si is typically deposited via LPCVD at temperatures >600°C. Furthermore, high-temperature (>900°C) annealing [21], [23] is typically required to achieve low stress. This imposes a severe constraint on the choice of contact electrode materials and does not meet the CMOS compatibility requirement. Nevertheless, phosphorus-doped n-type poly-Si is initially employed to demonstrate relay functionality. It can be deposited by LPCVD at 610°C and 375 mTorr process pressure with gas flow rates as follows: SiH_4 = 100 sccm, PH_3 = 4 sccm. Deposition rate is ~2nm/min.

Considering their high electrical conductivity and low thermal process budget, sputter-deposited metal films were considered to be reasonable candidates. A previous study investigating Al, Ni, Ti, TiN show that sputtered metals have very high residual stress, ranging from hundreds of MPa to 1 GPa [24]. Only Al shows low tensile residual stress

within the acceptable range (~10 MPa) but is hampered by very large strain gradient (~3×10^{-3} μm^{-1}) [24]. It was therefore decided not to pursue a metallic structural material.

Polycrystalline silicon germanium (poly-$Si_{1-x}Ge_x$, $x\leq0.6$) has been recently studied as low thermal budget structural material for MEMS [25]. It can be deposited by LPCVD at temperatures much lower than poly-Si (<450°C). While it tends to have larger strain gradient than poly-Si, it is still within an acceptable range when carefully optimized (~1×10^{-4} μm^{-1}). For this process, a very thin layer of amorphous silicon (<5 nm) is first deposited as a seed layer at 410°C and 300 mTorr, with Si_2H_6 = 100 sccm. Then, *in-situ*-boron-doped p-type poly-$Si_{0.4}Ge_{0.6}$ is deposited by LPCVD at 410°C and 600 mTorr with gas flow rates as follows: SiH_4 = 140 sccm, GeH_6 = 60 sccm, BCl_3 = 45 sccm. Deposition rate is ~6 nm/min. Poly-$Si_{0.4}Ge_{0.6}$ provides a low thermal budget alternative for the structural material in this work.

2.5 Process Integration Challenges

2.5.1 Film Delamination

In structures with moving parts, good adhesion between the layers is required. In this work, tungsten adhesion to SiO_2 proved to be an integration challenge. An assessment of film adhesion can quickly be determined by a scratch test or peel test [26]. The W-SiO_2 surface interaction energy is actually rather strong, but delamination occurs due to stress [27], [28], which is exacerbated by exposure to high temperatures. As described earlier, metal films tend to have high residual stress and strain gradient [24]. Residual stress consists of two components: the intrinsic stress of the film that depends on the microstructure of the film (grain size, geometry and crystal orientation) [29] and thermal stress that is due to difference in thermal expansion coefficients between the film and the underlying substrate when deposited or annealed at high temperatures. Tungsten residual stress depends strongly on the thickness of the film and deposition pressure during sputtering [30].

Tungsten is found to have a stronger adhesion to silicon. Therefore a layer of amorphous silicon (40 nm) is used as an adhesion layer between tungsten and SiO_2 to significantly increase the process temperature limit. Amorphous silicon is suitable because of its low deposition temperature compared to poly-Si and it is a smoother film due to the absence of microcrystalline grain structure. Amorphous silicon can be deposited via LPCVD at 530°C and 300 mTorr with gas flow rate SiH_4 = 200 sccm. Deposition rate is

~1.25nm/min. Even with an amorphous silicon adhesion layer, tungsten delamination will still occur at high temperatures. An 80 nm tungsten film survives 550°C anneal but delaminates when annealed at 610°C for ~2 hours. A 50 nm tungsten film survives 610°C anneal for <3 hours. Poly-Si deposited at 610°C tends to have compressive residual stress. Achieving low-stress poly-Si film requires annealing steps at temperatures >900°C. A rapid thermal anneal (RTA) at 950°C for 7 min results in 25 MPa tensile residual stress with negligible strain gradient [23]. It is also know that tungsten on silicon forms silicide (WSi_2) with low resistivity when subjected to rapid thermal annealing at temperatures >1000°C [31]. Rapid thermal annealing experiments are performed at 950°C, 1000°C and 1100°C for 30 sec, 1 min, 3 min and 6 min on a patterned tungsten film on SiO_2 with an A-Si adhesion layer. Delamination is observed in all cases due to stress.

The eventual switch to Al_2O_3 gate dielectric eliminated the need for an amorphous-silicon adhesion layer because tungsten adheres quite well to Al_2O_3. Nonetheless, processing temperatures still need to be kept below 550°C to avoid stress-induced delamination. The choice of tungsten as the electrode material imposes a strict process temperature limit.

2.5.2 Etching

Dry etching can be anisotropic, to achieve patterned features with better control and resolution than wet etching. In this work, etch recipes are developed for tungsten, Al_2O_3, SiO_2, A-Si, poly-Si, poly-SiGe. In developing an etch process, it is important to consider selectivity with respect to the underlying layer, especially when high precision is desired. For example, poor etch selectivity results in source/drain electrode thinning when patterning the sacrificial SiO_2 to form contact dimples, can undesirably increase the step height of the source/drain electrodes, and result in the formation of a gate dielectric "foot" around the periphery of the channel preventing it from ever contacting the source/drain electrodes. The etch rate also needs to be tuned depending on the thickness of the film. In general, it is recommended to complete the etch step between ~30 sec-90 sec. When the etch time is <30 sec (*i.e.* the etch rate is too fast), the etch process is less controllable and can easily result in too much overetch into the underlying layer. When the etch time is >90 sec (*i.e.* the etch rate is too slow), there can be issues with the photoresist mask. The substrate may overheat and burn the photoresist when it is etched continuously for >90 sec. If such long etch time is needed, it is recommended to separate the etch step into multiple cycles to allow the substrate to cool down between cycles.

Tungsten etch process is first developed in a Lam Research Reactive Ion Etch (RIE) system using CF_4 and O_2 as process gases. For small features such as the channel region, resist erosion during RIE etching is a concern due to the presence of O_2 in the plasma. Scanning electron micrograph (SEM) images of minimum sized channel (0.6 μm × 2.5 μm)

shows sloping edges due to resist erosion, in Figures 2.7(a) and 2.7(b). A set of experiments is conducted by varying CF_4/O_2 ratio while keeping the total flow rate at 100 sccm, RF power = 400 W, and pressure = 1 Torr (Figure 2.8). It is found that O_2 is required for etching to take place, and etch rate increases with increasing O_2 percentage. Devices with channels etched with CF_4/10% O_2 show lower yield (more prone to open circuit failure), possibly due to thinned and weakened channels that possibly leads to cracks, particularly where there is underlying edge topography (Figure 2.7(a)). The optimum etch recipe for channel patterning is obtained with CF_4/5% O_2 chemistry (Figure 2.7(b)), which shows the least amount of resist erosion and results in higher device yield.

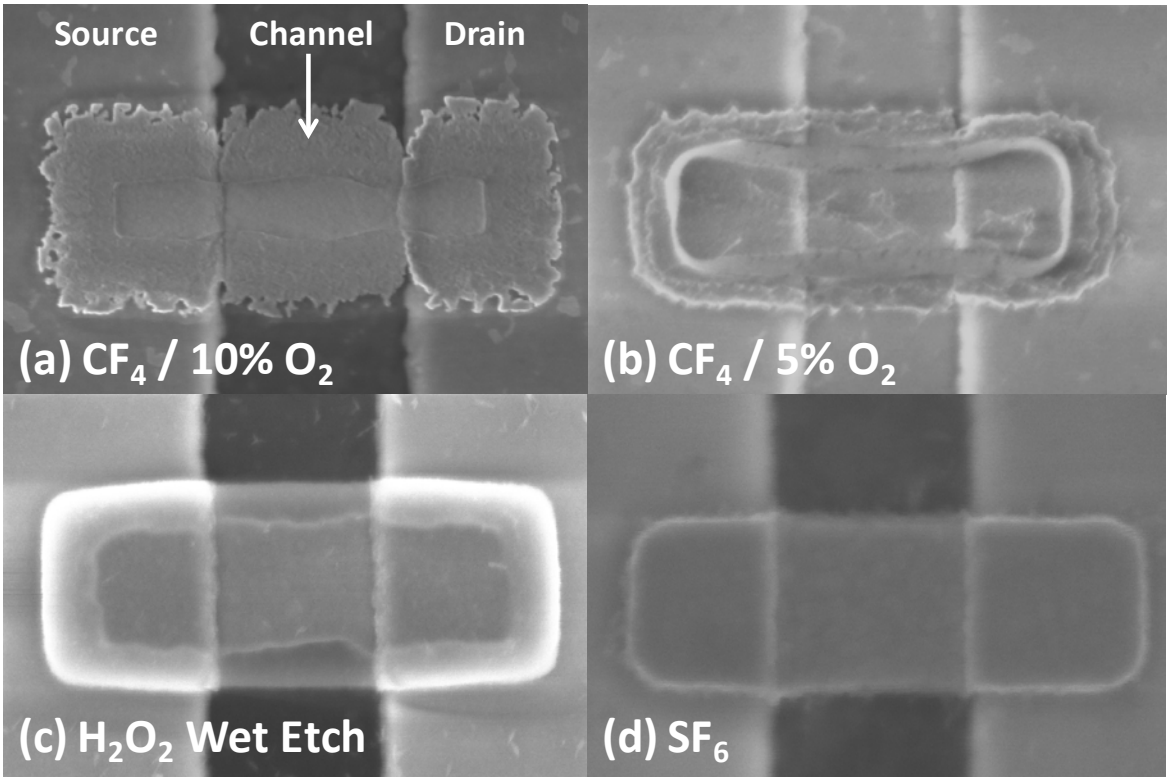


Figure 2.7: Scanning electron micrograph (SEM) images of tungsten channel after etching for minimum sized channel (0.6 µm × 2.5 µm), comparing different etching processes. (a) RIE with CF_4/ 10% O_2 chemistry. (b) RIE with CF_4/ 5% O_2 chemistry. (c) Wet etch with H_2O_2 for 17.5 min with A-Si hard mask. (d) RIE with SF_6 chemistry.

The effects of O_2 percentage on CF_4/O_2 and SF_6/O_2 processes have been previously studied in [32]. It is found that the etch rate initially increases with increasing O_2 percentage, as O_2 consumes fluorocarbon radicals and releases additional fluorine species for increased reaction rate to form tungsten hexafluoride (WF_6). However, the etch rate reaches a maximum at 20% O_2 content, beyond which it starts to drop. At high

concentrations of O_2, tungsten oxyfluoride (WOF_x) can be formed, which is less volatile than WF_6. WOF_x can form a barrier layer on the surface to impede the etching process.

CF_4 Flow Rate	O_2 Flow Rate	W ETCH RATES
100 sccm	0 sccm	Negligible
99 sccm	1 sccm	0.38 nm/sec
95 sccm	5 sccm	0.80 nm/sec
90 sccm	10 sccm	1.06 nm/sec
80 sccm	20 sccm	1.33 nm/sec

Figure 2.8: Tungsten etch rate in Lam Research RIE etcher with CF_4/O_2 etch chemistry at pressure = 1 Torr and RF power = 400 W. Total gas flow rate is kept at 100 sccm, while the gas composition is varied.

A wet etch process using H_2O_2 is investigated as an alternative. A-Si is used as a hard mask since photoresist was found to be insufficient to protect the features. A well-defined channel with vertical sidewalls is obtained using the wet etching process, but channel size is reduced due to lateral etching (Figure 2.7(c)). Also, the etch rate is very uncontrollable and non-uniform. Etching for a 50 nm film is completed anywhere between 10 min to 20 min, and etch completion can vary by a few minutes across different parts of the wafer. Endpoint is determined visually by detecting a color change. Etch residue is visually observed on the substrate even after rinsing. Nonetheless, functional devices are still found.

The arrival of an Applied Materials Centura Decoupled Plasma Source (DPS) metal etch chamber provided an improved dry etch process using SF_6 etch chemistry. It is found that the W etch rate is fastest in pure SF_6 gas, and decreases with higher O_2 percentage in a SF_6/O_2 mixture [32]. Therefore, O_2 is no longer used in this recipe to minimize photoresist erosion. Tungsten is etched with the following recipe: SF_6 = 85 sccm, pressure = 10 mTorr, source peak power = 1000 W, and bias peak power = 100 W. Note the lower process pressure as compared to the Lam Research RIE etcher. A more anisotropic etch due to more directional ion bombardment as a result of reduced scattering can be expected. Increased etch rate can be expected due to faster removal of etch by-products. Indeed, vertical sidewalls are obtained (Figure 2.7(d)) and etch rate is faster. The etch rate for various materials is given in Figure 2.9. Note that this recipe also has excellent W selectivity to Al_2O_3 (~30:1), which is useful for forming the source/drain electrodes.

An Al_2O_3 etch recipe is also developed in the Centura DPS chamber. The gas flow rates are BCl_3 = 45 sccm and Cl_2 = 90 sccm, with pressure=10 mTorr, source peak power = 1000 W, and bias peak power = 100 W. For all cases where Al_2O_3 needs to be etched, the etch stop layer is SiO_2. This recipe etches SiO_2 ~1.3× faster than Al_2O_3, which is not ideal.

Nonetheless, since the SiO_2 underneath the Al_2O_3 gate dielectric is sacrificial material (which eventually will be removed), this process is sufficient for our purpose.

FILMS	ETCH RATES	
	W Etch Recipe	***Al_2O_3 Etch Recipe***
	SF_6 = 85sccm	BCl_3 = 45sccm
		Cl_2 = 90sccm
	Pressure = 10mTorr	Pressure = 10mTorr
	Bias Peak Power = 100W	Bias Peak Power = 100W
	Source Peak Power = 1000W	Source Peak Power = 1000W
Sputtered W	2.0 nm/sec	1.14 nm/sec
Thermal Wet Oxide	2.7 nm/sec	2 nm/sec
LTO	3 nm/sec	2 nm/sec
LPCVD A-Si	14 nm/sec	5 nm/sec
LPCVD n^+poly-Si	30 nm/sec	5 nm/sec
Si substrate (bare wafer)	33.3 nm/sec	5.2 nm/sec
ALD Al_2O_3	0.068 nm/sec	1.5 nm/sec

Figure 2.9: Etch rates for W and Al_2O_3 etch recipes developed with the Applied Materials Centura DPS metal etcher. Etch rates are given for various common materials in this process to determine etch selectivity.

SiO_2, is etched in the Applied Materials Centura Magnetically Enhanced Reactive Ion Etch (RIE) chamber using the following recipe: CF_4 = 20 sccm, CHF_3 = 15 sccm, Ar =110 sccm, O_2 = 8 sccm, pressure = 50 mTorr, RF power = 500W. Etch rate is ~3.67 nm/sec. For the robust process presented in Chapters 3 and 4, an SiO_2 etch is needed to form the contact dimples, in which case selectivity to the underlying W electrode is important. For this recipe, selectivity of SiO_2 to W is ~1.8:1.

A-Si, poly-Si, and poly-SiGe are all etched using the standard silicon etch recipe in the Lam Research Transformer Coupled Plasma (TCP) etcher: HBr = 150 sccm, Cl_2 = 50 sccm, pressure = 12mTorr, TCP RF power = 300 W, Bias RF power = 150 W. Etch rate is found to be ~7 nm/sec. For a structural layer ~1 μm thick (involving a long etch time), an oxide hard mask is required to guarantee vertical sidewalls.

2.5.3 Gate Leakage & Gate Short Current

Despite functionality, early device prototypes are hampered by large gate leakage currents to the source and drain electrodes. Using ultra-thin (<10 nm) LTO as the gate dielectric leads to high gate leakage current. LTO is known to be leakier than thermal

oxide or PECVD TEOS oxides. LTO deposited at 400°C is measured to have a leakage current of ~100 nA/cm^2 for at field strength of 2 MV/cm, which is rather high. Additionally, due to some lateral etching during the HF vapor release step, LTO was removed at the edges of the channel and hence does not insulate the entire channel region. Within that lateral etch distance, the gate and channel are separated by an air gap. Surface leakage along the sidewalls of the LTO gate dielectric due to humidity [33] could lead to increased gate leakage current. Gate leakage is greatly reduced by replacing LTO with Al_2O_3 as the gate dielectric.

A second source for gate to source/drain current is lateral actuation of the beam caused by parasitic capacitances between the gate and the source/drain electrodes (Figure 2.10). In the early prototypes, gate dielectric only covers the region above the channel. The gate electrode, however, is the entire moving structure. A short circuit readily occurs if any unprotected part of the gate touches another electrode. During normal operation, the gate is at a much larger voltage (actuation voltage) than the source/drain electrodes, which are biased at low voltage. Therefore, lateral actuation forces toward the source/drain electrodes can exist. Applying a large drain bias while keeping the gate biased at 0V could cause the beam to bend sideways and stick to the drain (Figure 2.10(c)).

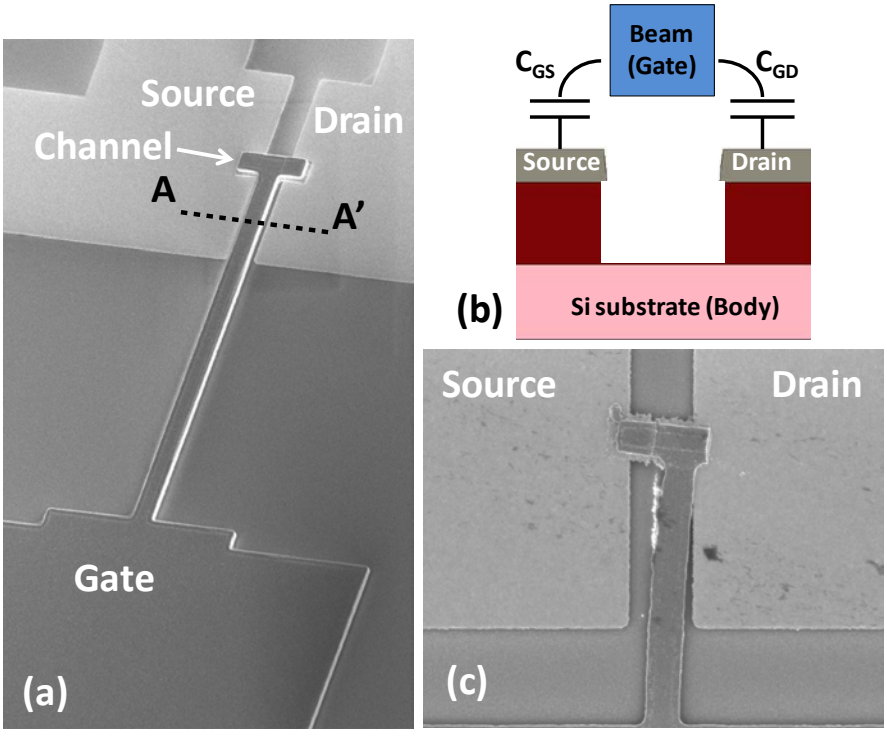


Figure 2.10: Lateral actuation due to parasitic capacitances between gate and source/drain. **(a)** Relay SEM image. **(b)** Illustration of parasitic capacitances at cross section A-A'. **(c)** A relay stuck to the drain after a high V_D is applied while V_G is at 0V.

Gate leakage current to the body is observed when a very high gate voltage is applied. This can be attributed to "catastrophic" pull-in where large parts of the structure collapse onto the substrate. This effect is especially common on buckled clamped-clamped beams. Very large applied voltage (>15V) is required to pull-in such buckled beams. When they do pull in, the entire structure tends to snap onto the substrate. Again, since most parts of the beam are not protected by the gate dielectric, large current flows from the gate to body. Due to the large actuation voltages involved (*i.e.* hence large force and current flow upon pull-in), this kind of shorting typically leads to device failure (fracture or permanent stiction).

It is therefore recommended to cover the bottom of the entire structure with the gate dielectric to prevent any possible electrical shorting. With Al_2O_3 gate dielectric, this is not a difficult fix. The gate dielectric simply can be patterned together with the structure. This is a self-aligned process and actually saves a lithography step. With this implemented, the probability of failure is greatly reduced since large gate current surges are no longer experienced.

2.5.4 Gate Dielectric "Foot"

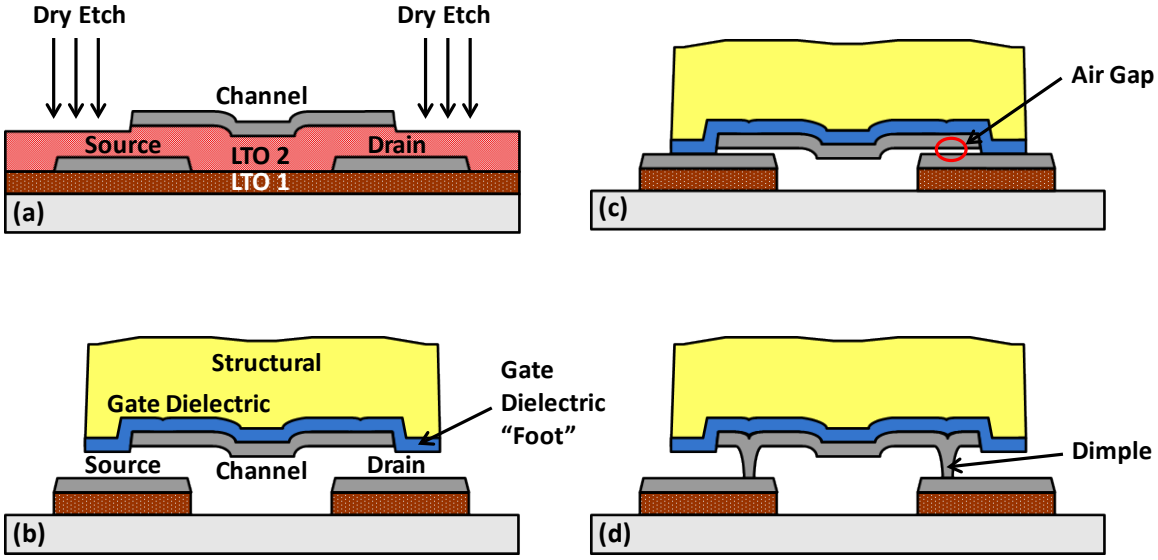


Figure 2.11: Illustration of gate dielectric "foot" problem. **(a)** Overetch during etching of channel. **(b)** Gate dielectric forms "foot" structure around the edges of the channel. **(c)** The "foot" structure prevents the channel and source/drain from making contact, leaving an air gap in between. **(d)** The use of contact dimples solves this problem.

At the end of every etch step, an overetch is always performed to ensure that the film being etched is completely removed across the wafer. This is especially important when there is an appreciable amount of non-uniformity in etch rate and/or film thickness across the wafer. In a non-production facility, such as a research or university lab, etching and deposition tools are constantly exposed to different conditions and materials, making it difficult to maintain uniformity. Since larger overetch may be required, good etch selectivity between the layer being etched and the etch stop (underlying) layer is desired, but not always available.

During channel formation, tungsten is etched with SiO_2 (LTO) as the etch stop layer (Figure 2.11(a)). From Figure 2.9, LTO is etched ~1.5× faster than W, so precise control is a challenge. Typically ~10-15 nm of the field LTO layer is etched away in order to complete the tungsten channel etch step. This recessed field causes the Al_2O_3 gate dielectric layer that is subsequently deposited to extend below the channel. When patterned, this Al_2O_3 layer forms a "foot" structure along the edges of the channel (Figure 2.11(c)) that eventually prevents the channel and source/drain from making contact. An air gap the size of the overetch depth exists between the channel and the source/drain after pull-in has occurred, as shown in Figure 2.11(c). Therefore, the relay fails to turn on.

Since overetch is unavoidable, a solution would be to add another lithography step to form a dimple at the contact region (Figure 2.11(d)). This additional processing step not only prevents this dielectric "foot" problem but also make the contact area well-defined, which is desirable to get consistent and controllable contact properties.

2.5.5 Stiction

Relays with longer cantilever beams (>15 μm) are found to be prone to stiction. Functional devices show poor endurance (<10 cycles). Two types of stiction are observed: 1) stiction during HF vapor release process, and 2) stiction during operation.

Stiction during release is caused primarily by surface adhesive forces. These include capillary force, van der Waals force, Casimir force, electrostatic force, and hydrogen-bond force [27], [34], [35]. Out of these forces, capillary force is typically dominant during release. Despite using a dry release process (HF vapor), the system is still humid since water is an etch by-product. In fact, liquid condensation on the electrostatic chuck in the HF vapor etch system is seen. Increasing the chuck temperature from room temperature to 40°C or 50°C to drive away water vapor and minimize condensation is found to help reduce stiction, albeit at the cost of a reduction in etch rate.

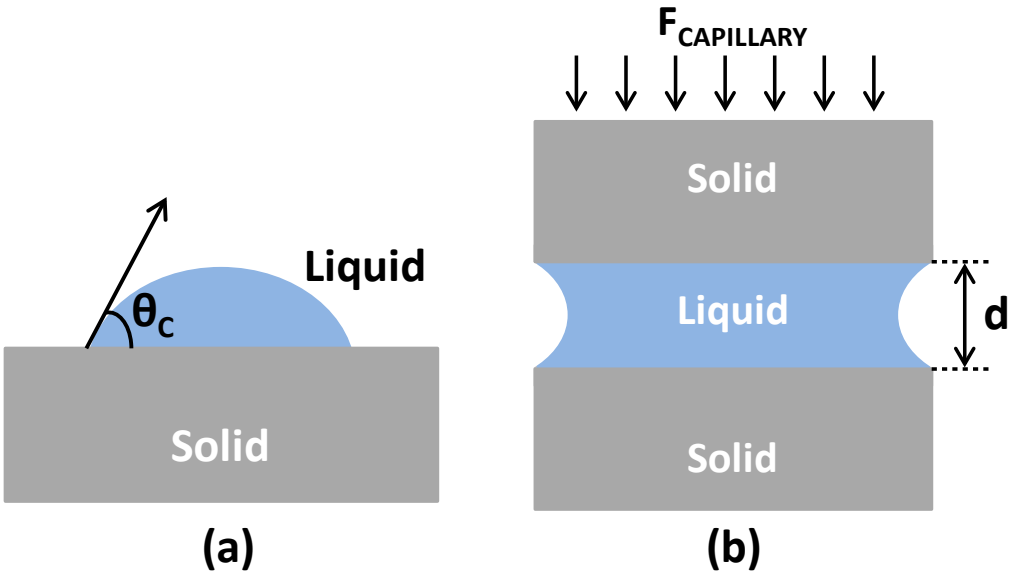


Figure 2.12: **(a)** The solid liquid interface. The contact angle (θ_C) determines wettability of the surface. **(b)** Illustration of capillary force when liquid is present between two solid plates.

The contact angle (θ_C) is a material property that determines the wettability of a surface, and thus the susceptibility of that surface to stiction due to capillary effect. When θ_C <90°, the surface is hydrophilic, while θ_C >90° indicates a hydrophobic surface. The surface interaction energy due to capillary forces (E_{CAP}) is given by [34]:

$$E_{CAP} = 2\gamma_l \cos\theta_C \quad \text{for} \quad d \leq d_{CAP} \tag{2.5.5.1}$$

$$E_{CAP} = 0 \quad \text{for} \quad d > d_{CAP} \tag{2.5.5.2}$$

where γ_l is the surface tension of the liquid (usually water) and θ_C is the contact angle of water on the surface.

Capillary condensation of water occurs when the gap between the solid plates (d) is smaller than the characteristic distance for capillary condensation (d_{CAP}) given by:

$$d_{CAP} = \frac{2\gamma_l \upsilon \cos\theta_C}{RT\log(RH)} \tag{2.5.5.3}$$

where υ is the liquid molar volume, R is the universal gas constant, T is the absolute temperature, and RH is relative humidity.

Stiction during operation is primarily caused by microwelding, due to Joule heating at the contacts when a high level of current flows. The voltage applied between the drain

and source electrodes of the relay (V_{DS}) needs to be carefully optimized to prevent microwelding. The 4T relay design provides the ability to optimize V_{DS} independent of the actuation voltage. However, in some of these relays a large gate current due to problems described in Section 2.5.3 makes the relays more prone to welding and needs to be solved.

One possible solution to mitigate stiction is to cover the contacts with a layer of ultra-thin TiO_2 coating [36]. TiO_2 may help to alleviate both stiction during release and stiction during operation. First, TiO_2 helps make the contacting surface less hydrophilic (θ_{CTiO2} >80° [37]) to reduce capillary forces. Tungsten readily forms native oxide (WO_3) at its surface in air (θ_{CWO3}<10° [38], [39]).

Second, TiO_2 helps limit the amount of current flow to the contact by making it slightly more resistive than a pure tungsten contact to reduce microwelding. Additionally, TiO_2 improves contact stability by serving as an oxidation barrier to slow down the formation of tungsten native oxide. Unlike WO_3, TiO_2 forms a relatively low potential barrier (~0.8 eV) for electrons to flow from W to TiO_2 to keep contact resistance within the acceptable range (<100 kΩ).

Highly conformal, ultra-thin TiO_2 can be deposited with atomic level control and precision through the PICOSUN SUNALE R150 Atomic Layer Deposition (ALD) tool. Integrating TiO_2 into the process is simple. As a protective layer post-release, it can be deposited after all other processing and release steps are done, since ALD is extremely conformal. Deposition is at 275°C, which comfortably satisfies the thermal budget requirement. If TiO_2 is to be used to protect the contact during release, then it has to be integrated during processing after electrode patterning and before channel deposition. TiO_2 coating is later incorporated into the process presented in Chapter 3 as an optional post-release step.

2.5.6 Structural Warping

Structural warping (out-of-plane deflection) is caused by residual stress and strain gradient. Residual stress represents the average stress in the film due to intrinsic stress related to the film microstructure and thermally induced stress from high-temperature treatments. Tensile films want to contract while compressive films want to expand to relieve the stress and reach equilibrium. Strain gradient represents changes in stress through the thickness of the beam. Per usual convention, positive (+) stress represents tensile stress while negative (-) represents compressive stress. A positive gradient indicates that the film is more tensile towards the top. In a cantilever structure after release, positive strain gradient results in upward bending because the bottom of the film expands more than the top of the film. A negative strain gradient causes downward bending since the top of the film expands more than the bottom of the film.

Poly-Si deposited at 610°C in this work tends to have high compressive residual stress and positive strain gradient. Cantilever beams tend to bend upwards while clamped-clamped beams are buckled upward, both of which increase pull-in voltage. A high-temperature anneal (>900°C) is required to achieve poly-Si films with low tensile residual stress and low strain gradient. However, this is not possible due to the thermal budget constraint strictly imposed by the tungsten electrodes.

Poly-SiGe beams deposited at 410°C also have stress issues. Interferometer measurements indicate that poly-SiGe cantilever beams bend downwards and clamped-clamped beams are buckled. Residual stress is compressive while strain gradient is negative. Negative strain gradient reduces the relay actuation gap, which results in lower actuation voltage but makes the structure more prone to stiction.

Stress-induced structural warping causes variation in actuation voltage, reduced reliability, and failure. It is therefore crucial to design structures that are more robust to stress and strain gradient. Conventional cantilever and clamped-clamped beam designs, while simple and compact, are not suitable. A design more robust to residual stress and strain gradient is implemented in Chapter 3.

2.5.7 Structural Fracture

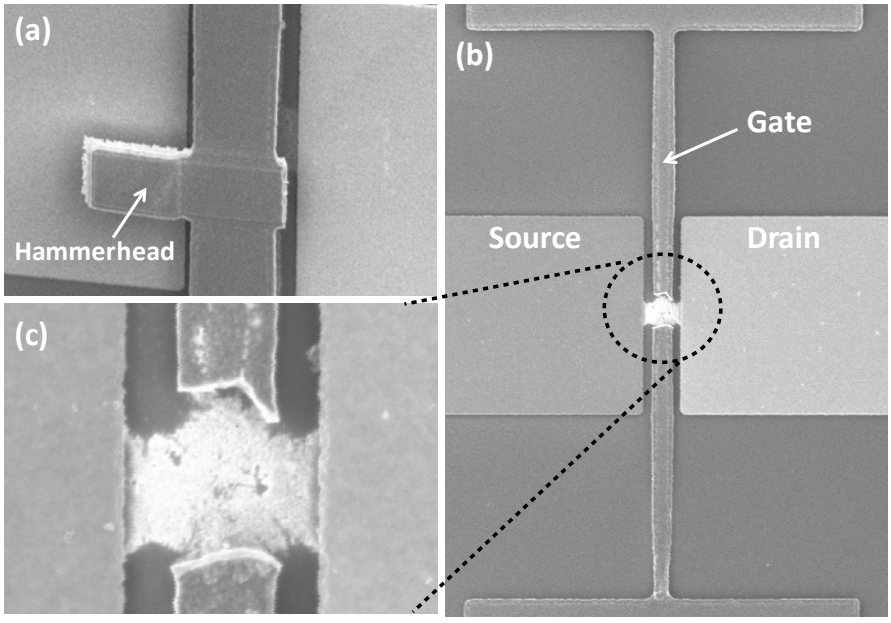


Figure 2.13: Fractures seen at the hammerhead after testing.

Poly-Si and poly-SiGe are strong structural materials for MEMS, with fracture strength between 1 GPa-3 GPa [40]-[42]. Structural failure of the beam is not observed under normal operation. Fracture is observed when relays are biased to very high voltages (>15 V), however, due to large electrostatic forces involved. SEM images of failed relays identify that the most common occurrence of fracture happens at the "hammerhead" where the channel is attached (Figure 2.13, 2.10(c)). In order to ensure reliability over billions of switching cycles, it is recommended to avoid small structural extension regions or designs that could potentially introduce weak spots, such as the hammerhead design.

2.6 References

[1] G. E. Moore, "Progress in digital integrated electronics," in Proc. International Electron Devices Meeting, pp. 11-15, 1975.

[2] W. W. Jang, J. O. Lee, J.-B. Yoon, M.-S. Kim, J.-M. Lee, S.-M. Kim, K.-H. Cho, D.-W. Kim, D. Park, and W.-S. Lee, "Fabrication and characterization of a nanoelectromechanical switch with 15-nm-thick suspension air gap," Applied Physics Letters, vol. 92, 103110, 2008.

[3] S. Chong, K. Akarvardar, R. Parsa, J.-B. Yoon, R. T. Howe, S. Mitra, H.-S. P. Wong, "Nanoelectromechanical (NEM) relays integrated with CMOS SRAM for improved stability and low leakage," in Proc. International Conference on Computer-Aided Design, pp. 478-484, 2009.

[4] J.-O. Lee, M.-W. Kim, S.-D. Ko, H.-O. Kang, W.-H. Bae, M.-H. Kang, K.-N. Kim, D.-E. Yoo, and J.-B. Yoon, "3-Terminal nanoelectromechanical switching device in insulating liquid media for low voltage operation and reliability improvement," in Proc. International Electron Devices Meeting, pp. 227-230, 2009.

[5] H. Kam, V. Pott, R. Nathanael, J. Jeon, E. Alon, and T.-J. King Liu, "Design and reliability of a micro-relay technology for zero-standby-power digital logic applications," in Proc. International Electron Devices Meeting, pp. 809-812, 2009.

[6] D. A. Czaplewski, G. A. Patrizi, G. M. Kraus, J. R. Wendt, C. D. Nordquist, S. L. Wolfley, M. S. Baker, and M. P. de Boer, "A nanomechanical switch for integration with CMOS logic," Journal of Micromechanics and Microengineering, vol. 19, 085003, 2009.

[7] G. K. Fedder, R. T. Howe, T.-J. King Liu, and E. Quevy, "Technologies for cofabricating MEMS and electronics," Proceedings of the IEEE, vol. 96, pp. 306-322, 2008.

[8] H. Takeuchi, A. Wun, X. Sun, R. T. Howe, and T.-J. King Liu, "Thermal budget limits of quarter-micrometer foundry CMOS for post-processing MEMS devices," IEEE Transactions on Electron Devices, vol. 52, pp. 2081-2086, 2005.

[9] P. M. Osterberg, S. D. Senturia, “M-TEST: A test chip for MEMS material property measurement using electrostatically actuated test structures,” Journal of Microelectromechanical Systems, vol. 6, no. 2, pp. 107-118, June 1997.
[10] P. J. Resnick and P. J. Clews, “Whole Wafer Critical Point Drying of MEMS Devices," Proc. SPIE 4558, Reliability, Testing, and Characterization of MEMS/MOEMS, pp. 189-196, Oct. 2001.
[11] R. Holm, Electric Contacts: Theory and Applications, Berlin, NY: Springer-Verlag, 1967.
[12] M. P. de Boer, J. A. Knapp, T. M. Mayer, and T. A. Michalske, “The role of interfacial properties on MEMS performance and reliability,” in Proc. SPIE, pp. 2-15, 1999.
[13] L. Kogut, and K. Komvopoulos, “Electrical contact resistance theory for conductive rough surfaces,” Journal of Applied Physics, vol. 94, pp. 3153-3162, 2003.
[14] G. M. Rebeiz, “RF MEMS: Theory, Design, and Technology,” New York: John Wiley & Sons, 2003.
[15] F. Chen, H. Kam, D. Markovic, T. King-Liu, V. Stojanovic, E. Alon, “Integrated Circuit Design with NEM Relays.” IEEE/ACM International Conference on Computer-Aided Design, pp 750-757, Nov. 2008.
[16] K. Williams, K. Gupta, and M. Wasilik, “Etch rates for micromachining processing - part II,” Journal of Microelectromechanical Systems, vol. 12, no. 6, pp. 761-778, Dec. 2003.
[17] D. Lee, H. Tran, and T.-J. King Liu, “Characterization of nanometerscale gap formation,” J. Electrochem. Soc., vol. 157, no. 1, pp. H94–H98, Nov. 2009.
[18] H. R. Shea, A. Gasparyan, H. B. Chan, S. Arney, R. E. Frahm, D. Lopez, S. Jin, R. P. McConnell, “Effects of electrical leakage currents on MEMS reliability and performance,” IEEE Transactions on Device and Materials Reliability, Vol. 4, No. 2, pp. 198-207, Jun. 2004.
[19] R. L. Puurunen, J. Saarilahti, and H. Kattelus, “Implementing ALD Layers in MEMS processing,” Electrochemical Society Transactions, vol. 11, no. 7, pp. 3-14, Oct. 2007.
[20] T. Bakke, J. Schmidt, M. Friedrichs, and B. Völker, “Etch Stop Materials for Release by Vapor HF Etching,” in Proceedings of the 16th Workshop on Workshop on Micromachining, Micromechanics, and Microsystems, pp. 103-106, Sept. 2005.
[21] R. T. Howe, B. E. Boser, and A. P. Pisano, “Polysilicon integrated microsystems: technologies and applications,” Sensors and Actuators A: Physical 56, no. 1, pp. 167-177, 1996.
[22] R. T. Howe and R. S. Muller, “Polycrystalline and amorphous silicon micromechanical beams: annealing and mechanical properties,” Sensors and Actuators 4, pp. 447-454, 1983.
[23] M. Biebl, G.T. Mulhem and R.T. Howe, “Low in sltu phosphorus doped polysilicon for integrated MEMS,” Tech. Digest, 8th Int, Conf.. Solid. State Sensors and Actuators (Transducers 95), Vol. I, pp. 198-201, Jun. 1995.

[24] J. Lai, “Novel Processes and Structures for Low Temperature Fabrication of Integrated Circuit Devices,” Ph.D. Dissertation, University of California, Berkeley, 2008.
[25] C. W. Low, T.-J. King Liu, and R. T. Howe, “Characterization of polycrystalline silicon-germanium film deposition for modularly integrated MEMS applications,” Journal of Microelectromechanical Systems, vol. 16, no. 1, pp. 68-77, Feb. 2007.
[26] B. N. Chapman, “Thin-film adhesion,” Journal of Vacuum Science and Technology, vol. 11, no. 1, pp. 106-113, 1974.
[27] V. K. Khanna, “Adhesion–delamination phenomena at the surfaces and interfaces in microelectronics and MEMS structures and packaged devices,” Journal of Physics D: Applied Physics, vol. 44, no. 3, 2011.
[28] R. H. Dauskardt, M. Lane, Q. Ma, and N. Krishna. "Adhesion and debonding of multi-layer thin film structures." Engineering Fracture Mechanics, vol. 61, no. 1, pp. 141-162, 1998.
[29] C. V. Thompson and R. Carel, “Stress and grain growth in thin films,” Journal of the Mechanics and Physics of Solids, vol. 44, no. 5, pp. 657-673, 1996.
[30] Y. G. Shen, Y. W. Mai, Q. C. Zhang, D. R. McKenzie, W. D. McFall, and W. E. McBride. "Residual stress, microstructure, and structure of tungsten thin films deposited by magnetron sputtering." Journal of Applied Physics vol. 87, no.1, pp. 177-187, 2000.
[31] M. P. Siegal, and J. J. Santiago, “Effects of rapid thermal processing on the formation of uniform tetragonal tungsten disilicide films on Si (100) substrates,” Journal of Applied Physics, vol. 63, no. 2, pp. 525-529, 1988.
[32] W. S. Pan and A. J. Steckl, “Selective reactive ion etching of tungsten films in CHF_3 and other fluorinated gases,” Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures vol. 6, no. 4, pp. 1073-1080, 1988.
[33] H. R. Shea, A. Gasparyan, H. B. Chan, S. Arney, R. E. Frahm, D. López, S. Jin, and R. P. McConnell, ‘Effects of electrical leakage currents on MEMS reliability and performance,” IEEE Transactions on Device and Materials Reliability, vol. 4, no. 2, pp. 198-207, Jun. 2004.
[34] W. M. van Spengen, R. Puers, and I. De Wolf. "A physical model to predict stiction in MEMS." Journal of micromechanics and microengineering, vol. 12, no. 5, pp. 702-713, 2002.
[35] D. T. Lee, “Novel Nanoscale Electromechanical Systems Devices and Technology,” Ph.D. Dissertation, University of California, Berkeley, 2009.
[36] V. Pott, H. Kam, J. Jeon, and T.-J. King Liu, “Improvement in mechanical contact reliability with ALD TiO_2 coating,” in Proc. AVS Conference, pp. 208-209, 2009.
[37] G. Triani, J. A. Campbell, P. J. Evans, J. Davis, B. A. Latella, and R. P. Burford, “Low temperature atomic layer deposition of titania thin films,” Thin Solid Films, vol. 518, no. 12, pp. 3182-3189, 2010.
[38] R. Azimirad, N. Naseri, O. Akhavan, and A. Z. Moshfegh, “Hydrophilicity variation of WO_3 thin films with annealing temperature,” Journal of Physics D: Applied Physics, vol. 40, no. 4, pp. 1134-1137, 2007.

[39] M. Miyauchi, A. Nakajima, T. Watanabe, and K. Hashimoto, “Photocatalysis and photoinduced hydrophilicity of various metal oxide thin films,” Chemistry of Materials, vol. 14, no. 6, pp. 2812-2816, 2002.
[40] J. Bagdahn, W. N. Sharpe Jr, and O. Jadaan, “Fracture strength of polysilicon at stress concentrations,” Journal of Microelectromechanical Systems, vol. 12, no. 3, pp. 302-312, 2003.
[41] H. Kapels, R. Aigner, and J. Binder, “Fracture strength and fatigue of polysilicon determined by a novel thermal actuator [MEMS],” IEEE Transactions on Electron Devices, vol. 47, no. 7, pp. 1522-1528, 2000.
[42] R. Modlinski, A. Witvrouw, A. Verbist, R. Puers, and I. De Wolf, “Mechanical characterization of poly-SiGe layers for CMOS–MEMS integrated application,” Journal of Micromechanics and Microengineering, vol. 20, no. 1, 2009.

Chapter 3

4-Terminal Relay Technology

3.1 Introduction

Nano-electro-mechanical (NEM) relay technology recently has been proposed for ultra-low-power digital integrated circuit applications [1], [2]. This is because the relay is an ideal switch, in that it exhibits abrupt on/off switching behavior and zero off-state leakage current (I_{OFF}), so that its operating voltage (V_{DD}) can be reduced to be close to zero, in principle. Thus, NEM relay technology can potentially overcome the fundamental energy efficiency limit of CMOS technology [3]. In order to realize this promise, however, relays must have high endurance and low switching voltages. The former requirement has been difficult to achieve because of surface wear and stiction-induced failure. The latter requirement is difficult to achieve with a conventional 3-Terminal (3T) relay design, particularly for highly scaled dimensions [1]. In order to achieve low voltage operation, low pull-in voltage is desired. This is difficult to achieve through layout and process. Longer beams (smaller spring constant) or larger actuation area (higher electrostatic force) can be employed at the expense of larger area. Reducing the air gap thickness potentially increase susceptibility to stiction during release. A thinner structural material with lower spring constant is often hampered by strain gradient problems. Furthermore, there will always be some variability in pull-in voltage due to process-induced variations.

In this chapter, a highly reliable mechanical contact technology employing tungsten electrodes is developed and a 4-Terminal (4T) relay technology is proposed to overcome these challenge for complementary logic circuit applications. The 4T relay design provides a convenient way of electrically adjusting the gate switching voltages post-process via body biasing for low-voltage operation. As a result, a 4T relay can mimic the operation of either an n-channel or p-channel MOSFET. Prototype relays fabricated with a CMOS-compatible process are demonstrated. Fabricated 4T relays exhibit good on-state current (I_{ON} > 700μA for V_{DS} = 1V) and zero off-state leakage current with subthreshold swing <0.1 mV/dec. Low-voltage switching (<2 V) and low switching delay (100 ns) are demonstrated by appropriately biasing the body terminal. Endurance exceeds 10^9 on/off cycles without stiction or wear issues. Therefore, the 4T relay technology is promising to realize relay-based integrated circuits.

3.2 Robust 1st Generation 4-Terminal Relay Structure

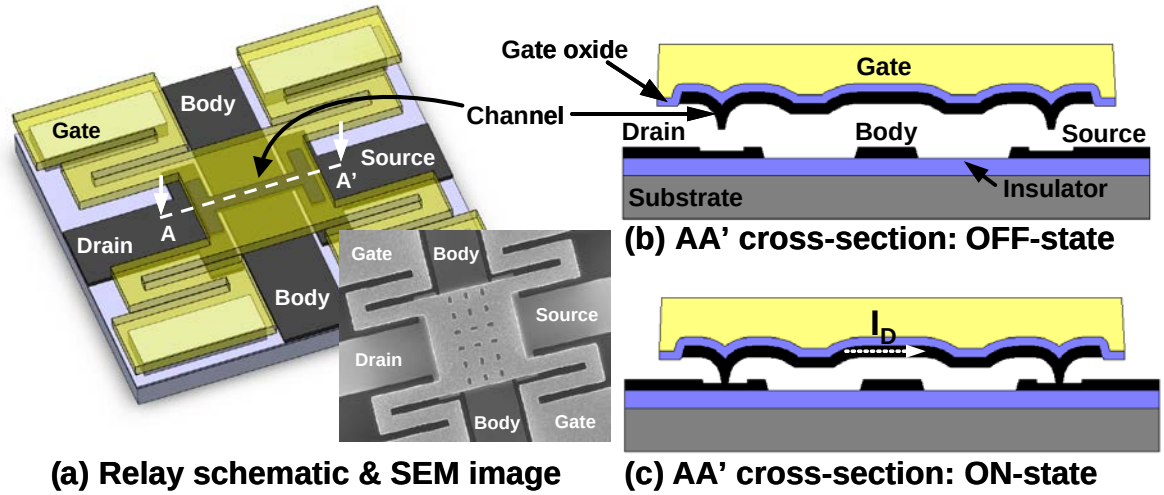


Figure 3.1: (a) Schematic illustrations of the four-terminal-relay structure. (a) Isometric view and plan-view SEM image. (b) Cross-sectional view along the channel (A-A') in the OFF state. (c) Cross-sectional view along the channel (A-A') in the ON state.

Figure 3.1 illustrates the 4T relay structure and its operation. It comprises a movable conductive plate, the gate electrode, suspended at its corners by four folded-flexure beams. A metallic conducting channel is attached via an insulating dielectric to the movable gate. In the off state, an air gap separates the channel from the underlying metallic source and drain electrodes, so that no current can flow. In the on state, electrostatic force between the gate and the underlying body electrode causes the gate to be actuated downward sufficiently to bring the channel into contact with the source and drain electrodes, to form a conductive path for current to flow. The relay switches on abruptly as the magnitude of the gate-to-body voltage ($|V_{GB}|$) is increased above the pull-in voltage (V_{PI}) and switches off abruptly as $|V_{GB}|$ is decreased below the release voltage (V_{RL}). The key dimensions of the structure are given in Figure 3.2.

This four-terminal relay design solves many of the problems that were encountered in the previous designs (Chapter 2). The key features of this design are as follows:

(1) Clamped-clamped structure.
The movable structure is essentially a clamped-clamped beam. Recall from Section 2.5.6 that a simple cantilever design tends to not have sufficient spring restoring force to ensure that the relay turns off reliably. While a clamped-clamped beam design provides the necessary restoring force, it tends to buckle upon release due to compressive residual stress. A structural material with zero

residual stress or slightly tensile (<100 MPa) is ideal, but difficult to achieve. Therefore, a folded-flexure suspension beam design is used to relieve residual stress in the structural layer and prevent buckling, while still taking advantage of the higher spring restoring force of a clamped-clamped beam for reliable turn-off. Moreover, the folded flexure makes this design more robust to thermal variations, as it helps mitigate any effect of thermal stress as well.

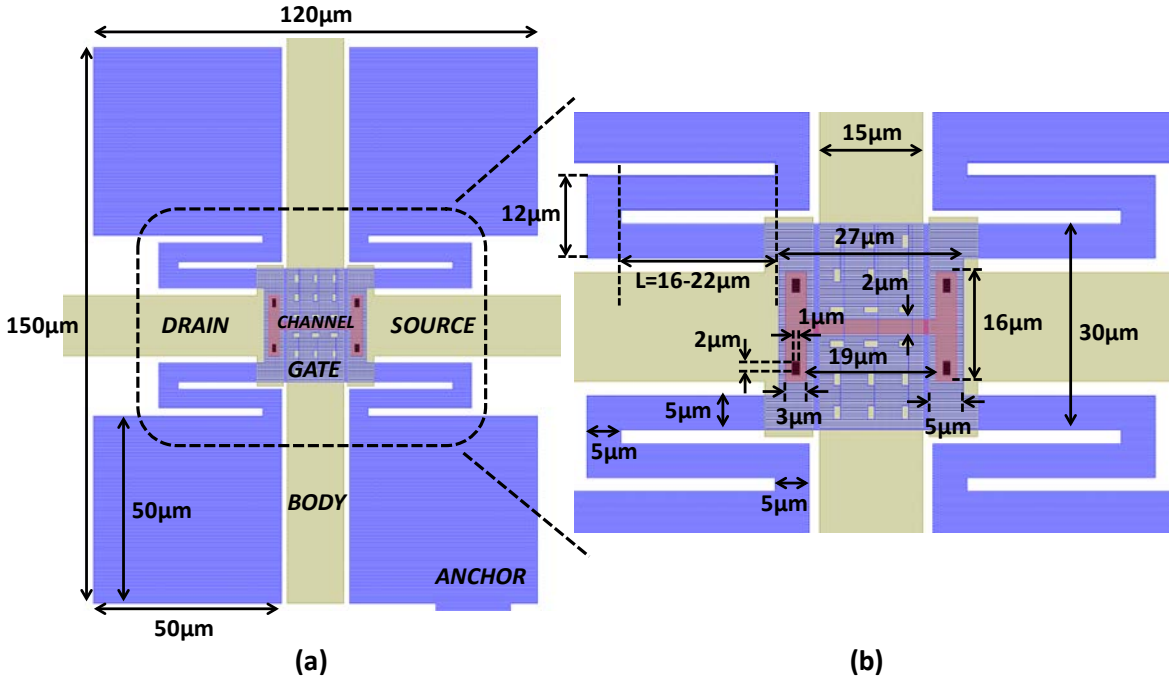


Figure 3.2: 1st generation 4T relay design. (a) Key dimensions. (b) Zoomed-in view showing channel and dimple regions. Total footprint area is 18000 µm^2 (120 µm × 150 µm).

(2) Four symmetric flexures.
Having four symmetric flexures instead of two (as in conventional clamped-clamped beam) gives extra stability and prevents rotation of the plate due to residual stress.

(3) Dimpled contacts
The dimple design allows the apparent contacting area to be well-defined by lithography. The use of dimpled contacts also prevents the gate dielectric "foot" issue due to overetch into the sacrificial layer when patterning the channel (Section 2.4.4).

(4) Actuation area and flexures are decoupled.
The plate at the center defines the actuation area. This way, the flexure dimensions (which determine the spring restoring force) can be optimized separately from the area that contributes to electrostatic force for actuation.

(5) Square plate with channel underneath the plate

For structural strength, the actuation plate is square without any small extensions such as a hammerhead that could introduce weak spots prone to fracture (Section 2.4.7). The channel lies underneath the plate itself.

(6) Gate oxide underneath the entire structural layer.
The presence of gate oxide underneath the entire structure prevents the possibility of electrical shorting to the gate electrode even if the structure collapses, which could happen due to structural warping induced by strain gradient, parasitic actuation by source/drain electrodes, and catastrophic pull-in (Section 2.4.3).

3.3 4-Mask Process

Following the lessons learned from the initial process development work described in Chapter 2, a robust 4-Terminal relay process flow (Figure 3.3) is developed, using the four-terminal relay design shown in Figure 3.1. This process has been successfully demonstrated in the UC Berkeley Marvell Nanofabrication Laboratory with excellent repeatability and yield (>95%).

Silicon dioxide (SiO_2) deposited at low temperature (400°C) by low-pressure chemical vapor deposition (LPCVD) was used as the sacrificial material, because it is easily selectively removed in HF vapor to release the gate stack structure. Note that the thickness of the SiO_2 in the source/drain contact regions (which determines the contact gap) is thinner than that of the SiO_2 over the body electrode (which determines the actuation gap), in order to reduce contact velocity [1] for better reliability. Specifically, the contact gap (100 nm) is one-half the actuation gap (200 nm) for the most energy-efficient operation [4]. Aluminum oxide (Al_2O_3) deposited at 300°C by atomic layer deposition (ALD) was used to protect the oxidized Si wafer substrate from the HF-vapor release etch, and also was used as the gate dielectric material. Tungsten (W) deposited by DC magnetron sputtering was used as the material for the source, drain, body, and channel electrodes because it is resistant to HF-vapor etching [6] and physical wear [7]. *In-situ* boron-doped polycrystalline silicon-germanium (poly-$Si_{0.4}Ge_{0.6}$) deposited at 410°C by LPCVD was used as a high-quality, low-thermal-budget structural material [8] for post-CMOS integration capability. Lastly, an ultra-thin (~3Å thick) coating of titanium dioxide (TiO_2) was deposited at 300°C by ALD after the gate-stack structure was released, to improve contact reliability [5].

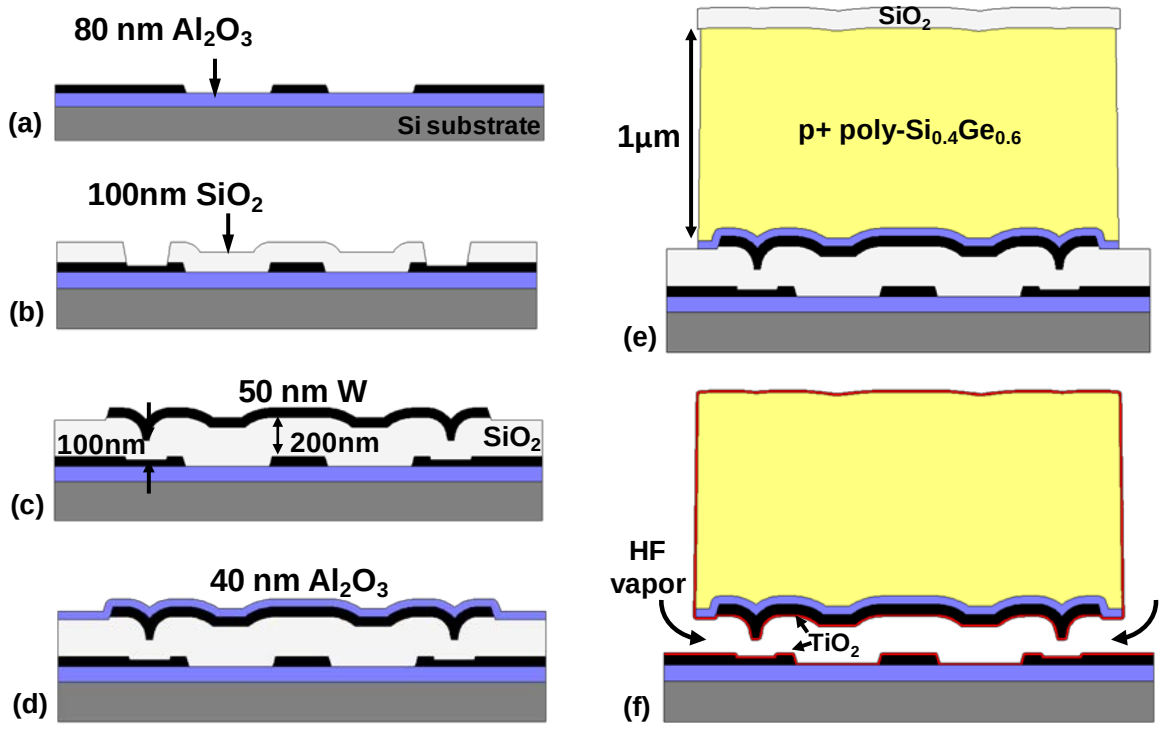


Figure 3.3: Illustration of the 4-mask process used to fabricate the 1st prototype four terminal relays, shown along cross section A-A' of Figure 3.1. (a) 80 nm Al_2O_3 layer deposited and 50 nm W layer deposited and patterned to form source, drain, and body electrodes. (b) 1st sacrificial SiO_2 layer (100 nm) deposited and source/drain contact regions defined. (c) 2nd sacrificial SiO_2 layer (100 nm) deposited and 50 nm W layer deposited and patterned to form channel. (d) 40 nm Al_2O_3 gate dielectric layer deposited. (e) 1 µm p^+ poly-$Si_{0.4}Ge_{0.6}$ deposited and patterned using LTO hard mask to form gate electrode. (f) Gate-stack released in HF vapor and coated with 3Å TiO_2.

3.4 Characterization Results

3.4.1 DC Characteristics

The 1st working 4T relay prototypes fabricated using the 4-Mask process are electrically characterized using the HP4156 semiconductor parameter analyzer. Measured I_{DS}-V_{DS} characteristics (Figure 3.4) are linear, which indicates that ohmic contacts are

achieved despite the TiO_2 electrode coatings. The on-state resistance, R_{ON}, is relatively insensitive to the gate overdrive (V_G-V_{PI}), as expected for a hard contact material. It should be noted that significant variations in on-state resistance, R_{ON}, are seen from device to device; however, this should not significantly impact the performance of an optimally designed relay circuit, because throughput will be limited by mechanical delay rather than by electrical delay [2].

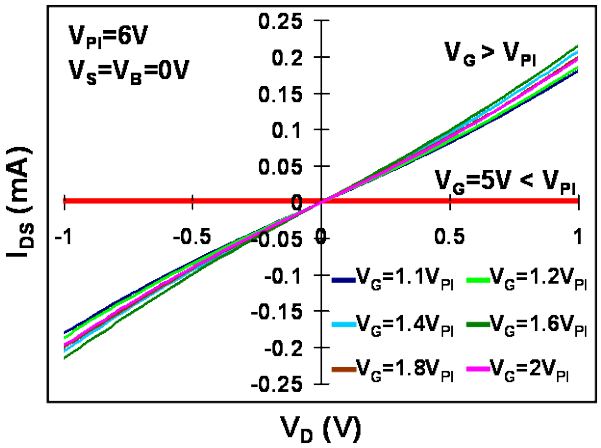


Figure 3.4: Measured I_{DS}-V_D for V_G<V_{PI} (off-state) and V_G>V_{PI} ranging from 1.1V_{PI} to 2V_{PI} (on state). Pull-in voltage V_{PI} = 6V, V_S = V_B = 0 V.

Measured I_{DS}-V_G characteristics (Figures 3.5, 3.6, 3.7) show zero I_{OFF} and relatively high on-state current (I_{ON} > 800 µA for V_{DS} = 1V). The ~10^{-14} A leakage seen for the off-state represents the noise level in the measurement setup. Hysteretic behavior ($V_{PI} \neq V_{RL}$) is seen in the I_{DS}-V_G curves because the relay is operating in pull-in mode [9], and also because there is non-zero surface adhesive force. Measurements with very fine voltage step size (1 mV) in Figure 3.5 show SS < 0.1mV/dec, confirming hyper-abrupt switching behavior. The ambipolar nature of electrostatic attraction allows the relay to be turned on with either a positive V_{GB} or a negative V_{GB}. This allows the 4T relay to operate mimicking either an n-channel or p-channel MOSFET, with appropriate biasing of the body electrode. As shown in Figure 3.6, N-relay operation (switching at V_G > 0 V) or P-relay operation (switching at V_G < 0 V) is achieved by applying a body bias (V_B) of 0 V or V_{DD}, respectively.

Since V_{PI} is dependent on relay design parameters (*e.g.* flexure beam dimensions, actuation area, and air-gap thicknesses) and can vary as a result of process-induced variations, the ability to tune V_{PI} post-process via V_B is an important advantage of the 4T relay design. Figure 3.7 shows how the switching voltages change with V_B. Since the maximum sweep voltage is kept the same for all V_B conditions, the hysteresis voltage increases as the maximum gate overdrive (V_G-V_{PI}) increases, due to gate-dielectric charging.

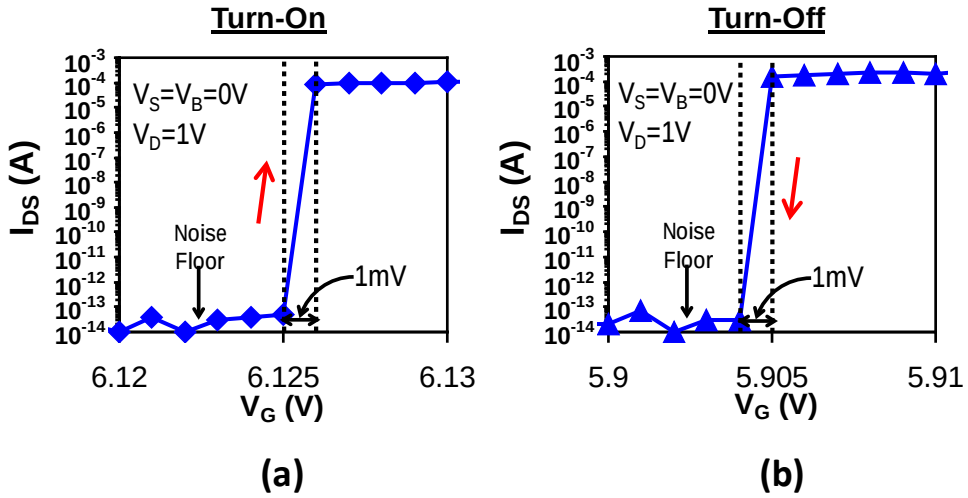


Figure 3.5: Measured I_{DS}-V_G characteristics of a single 4T relay, with V_D = 1 V and V_S = V_B = 0 V. Very fine (1 mV) voltage step size is used to examine the (a) turn-on and (b) turn-off characteristics. Note that I_{DS} in the off-state is at the noise floor of the measurement instrument and switching is hyper-abrupt, with >10 orders of magnitude jump in I_{DS} for a 1mV step (i.e. SS < 0.1 mV/dec).

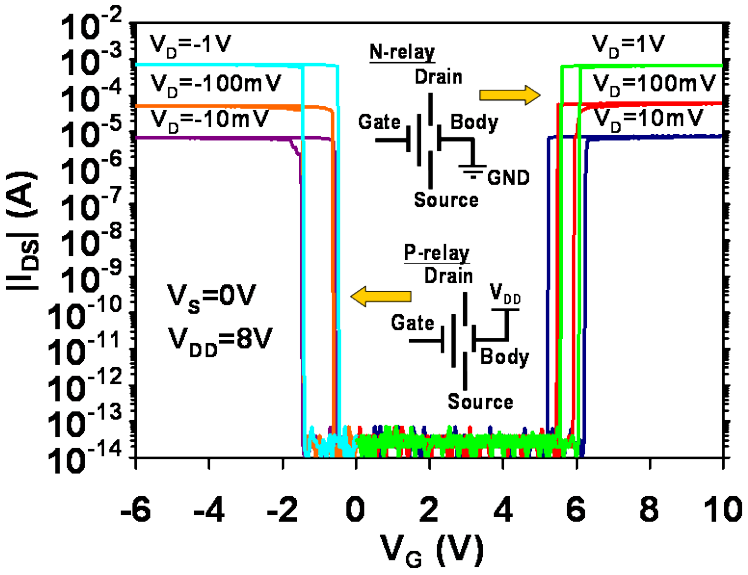


Figure 3.6: Measured I_{DS}-V_G characteristics of a single 4T relay, with V_D = 10mV, 100 mV, and 1 V and V_S = 0 V. Operation mimicking that of either an n-channel or a p-channel MOSFET is seen by biasing the body at either 0V or V_{DD}, respectively.

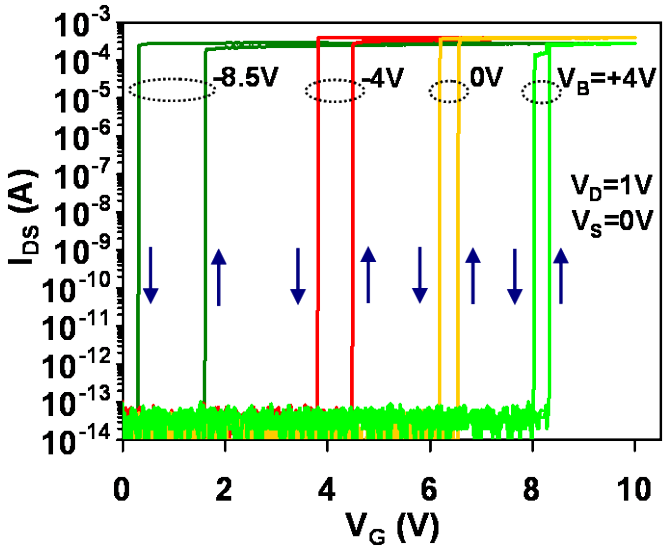
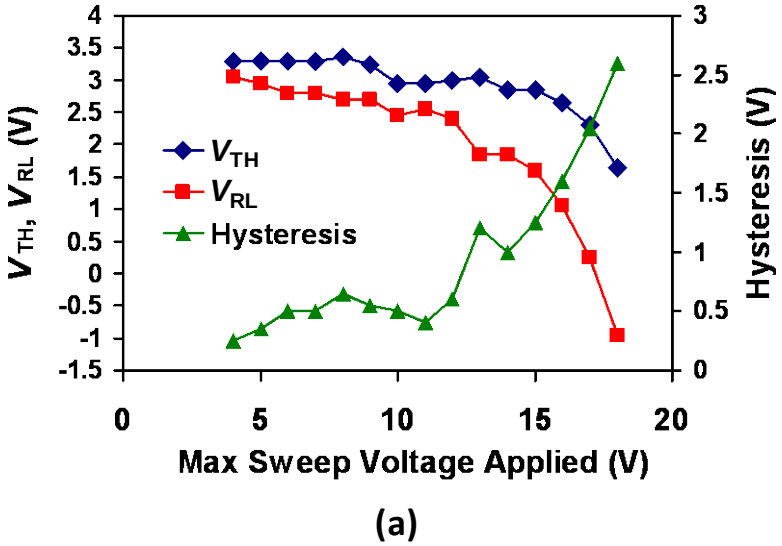


Figure 3.7: Measured I_{DS}-V_G characteristics of a single 4T relay, with V_D=1V, V_S=0V, and V_B=4V, 0V, -4V, and -8.5V. V_{PI} can be tuned by adjusting V_B.

A more thorough investigation of this charging effect is conducted by varying the maximum sweep voltage for successive measurements, while keeping V_B = 0 (Figure 3.8(a)). Measurements are done in the order from smallest to largest maximum sweep voltage. As the device is stressed with increasingly higher gate overdrive, decreasing V_{PI} and increasing hysteresis voltage are observed. This implies that holes (positive charges) are being injected into the gate dielectric as it becomes easier (*i.e.* requires lower voltage) to pull-in and release the relay with increasing charge.



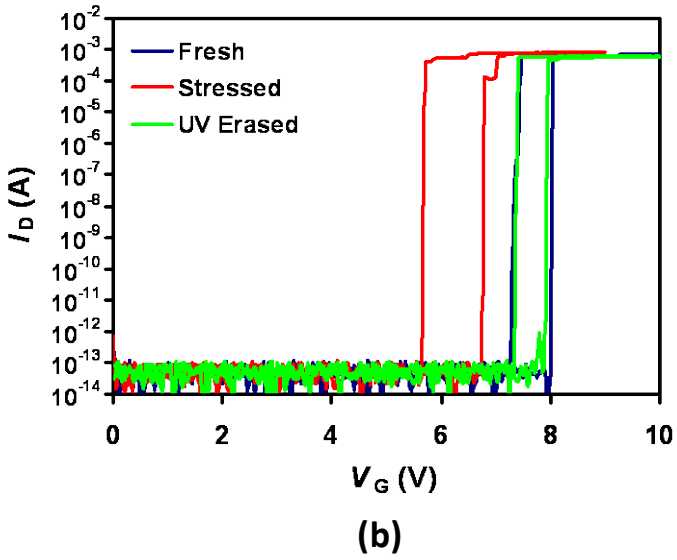


Figure 3.8: Dielectric charging phenomenon. **(a)** Measured V_{PI} and V_{RL} and hysteresis voltage with increasing gate overdrives. Measurements were taken successively from lowest maximum sweep voltage to the highest. **(b)** Measured I_{DS}-V_G characteristics of a single 4T relay, showing the effect of electrical stress (V_{GB}=2.5V_{PI} for 10min) on a fresh device and after exposure to UV light for 15min. The ability of UV erase to restore the original (pre-stress) switching voltages verifies that the switching voltage shifts seen are indeed due to a gate-oxide charging phenomenon.

In Figure 3.8(b), a fresh device is stressed at V_{GB} = 2.5V_{PI} for 10 min, then subjected to ultraviolet (UV) light for 15 minutes. The device is re-measured and the switching voltages returned to its pre-stress values, confirming that this effect is indeed a charging phenomenon. Gate dielectric charging allows the 4T relay to be potentially useful as a memory device if designed appropriately [10]. When used as a logic device, where charging is undesired, excessive gate overdrive should be avoided.

3.4.2 Switching Speed

Since the speed of relay ICs is limited by mechanical delay, as opposed to RC delay (associated with capacitive charging or discharging) typical in CMOS ICs [2], it is important to characterize the mechanical switching speed of the relay. The turn-on delay is expected to be much slower than the turn-off delay, because the relay needs to travel the distance of the contact gap (~100 nm in this particular relay) in order for contact to be made between the channel and the source/drain regions. On the other hand, a much smaller travel distance suffices in turning off the relay. Once the contact is broken and channel is lifted sufficiently so that tunneling current through air is negligible (~2 nm), the relay is off.

Therefore, the mechanical delay likely will be limited by the turn-on delay, which is characterized in Figures 3.9 and 3.10.

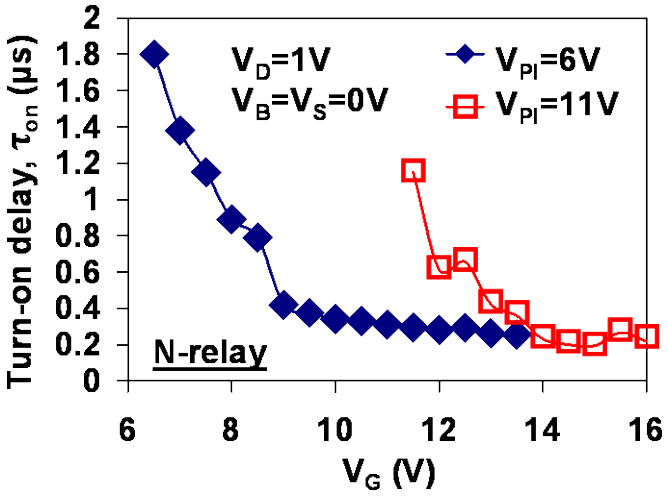


Figure 3.9: Measured turn-on delay (τ_{on}) with increasing applied gate voltage (V_G), for two relays with different flexure lengths (L). The two relays have V_{PI} = 6 V and V_{PI} = 11 V. V_D = 1 V and V_S = V_B = 0 V. Delay decreases with higher V_G, but eventually saturates at ~200 ns.

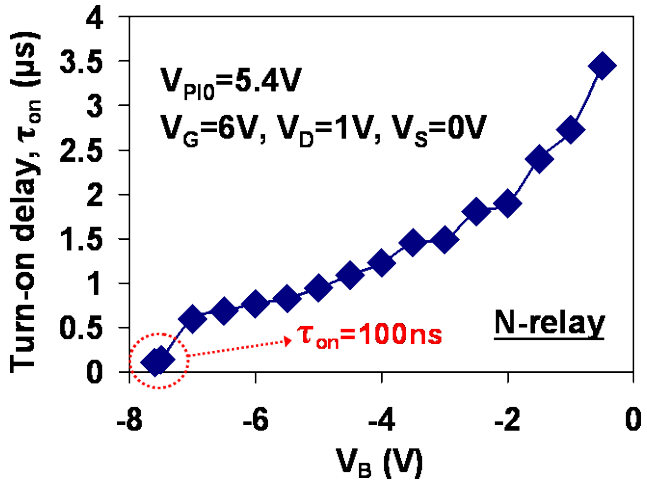


Figure 3.10: Turn-on delay (τ_{on}) *vs.* body bias (V_B). Relay has V_{PI0} = 5.4 V. V_G = 6 V, V_D = 1 V, and V_S = 0 V applied for all measurements. Body bias allows delay to be reduced

below the 200 ns minimum delay barrier (Fig. 9). With V_B = -7.6V, a τ_{on} = 100 ns is achievable.

As shown in Figure 3.9, the relay turn-on delay (τ_{on}) decreases with increasing gate overdrive, eventually reaching a limit of ~200 ns regardless of flexure length. A gate overdrive of ~2 V would be optimal to operate the relay at close to its maximum speed (since excessive gate overdrive leads to gate dielectric charging). This limit can be broken by applying a non-zero V_B to effectively reduce the contact gap. Since the relay is already pre-actuated down a certain distance prior to the actuation event, it needs to travel less distance to make contact compared to a case where V_B = 0 V. In this manner of operation, τ_{on} as low as 100 ns was achieved (Figure 3.10).

From these results, the switching delay of relays is in the 100 ns to1 μs range, which corresponds to circuit speeds in the range of 1 MHz - 10MHz. Modeling efforts have shown that the switching speed improves with scaling of the relay dimensions. For a 90 nm relay technology, switching delay is in the order of 10 ns (100 MHz operating frequency) [4].

3.4.3 Endurance

Endurance is of paramount importance for digital IC applications. Even more so than speed or power consumption, it is crucial that the system remains functional over its expected lifetime. The industry standard for device operation lifetime is 10 years. As a reasonable benchmark, in a relay-based microcontroller for embedded sensor applications operating at 100 MHz with 0.01 average transition probability, relays experience ~3×10^{14} on/off cycles over 10 years. Therefore, it is fair to say relays for digital ICs require endurance >10^{14} cycles.

CMOS reliability concerns include shifts in saturation drain current (I_{DSAT}) or threshold voltage (V_{TH}) due to phenomena such as hot carrier injection (HCI) and negative bias temperature instability (NBTI), or gate leakage due to time dependent dielectric breakdown (TDDB) [11]. In relays, the failure mechanism is typically associated with either the contact, structure, or dielectric [12]. In relays fabricated in this work, structural fracture or fatigue is not expected to be an issue since poly-SiGe is a structurally robust material [13]. Dielectric charging is observed (Section 3.4.1), but can be mitigated so long as gate overdrive is not excessive. The contact is the primary reliability concern. Contact endurance can be viewed from two different aspects: failure and performance degradation.

From a contact failure perspective, reliability issues include: 1) stiction due to microwelding and surface adhesive forces, and 2) wear and plastic deformation. The use of

tungsten (high hardness, high melting point) as contact material solves most of these concerns. The benefits of tungsten are discussed in depth in Section 2.3.2.

From a performance perspective, the contact determines relay on-resistance (ref. Section 2.3.2). For an optimal digital IC design using relays, R_{ON} can be 10-100 kΩ to guarantee that the electrical charging delay will be much smaller than the mechanical switching delay (100 ns), for typical load capacitances (10-100 fF) [2]. Intrinsically, tungsten used in this technology has sufficiently low contact resistance (~1 kΩ) to comfortably meet this requirement. However, hard materials are prone to chemical reaction as opposed to noble metals (*e.g.* gold). Tungsten readily oxidizes in air to form native oxide (WO_3) that significantly increases its contact resistance. Keeping a low contact resistance turns out to be quite a challenge.

In this work, an attempt is made to monitor the evolution of contact properties after each cycle. The relay is connected to the HP4156 parameter analyzer as for the DC measurements. The parameter analyzer is programmed to bias and take measurements successively for a specific number of cycles. Therefore, each time the relay turns on, I_{DS} is measured. By doing so, the evolution of the contact resistance with each cycle can be monitored at the expense of testing speed. Since testing is significantly slower at ~1 second per on/off cycle (1 Hz), only 10,000 cycles can be practically captured. The W native oxide at the contact is electrically broken before each measurement so that each device starts with the same initial R_{ON} (~1 kΩ). The initial native oxide break is done by switching the device on and applying high drain voltage (V_D = 4-5 V). The bias is automatically stopped when a drain current spike is detected, indicating native oxide breakdown. Adhesion forces of the 4T relays is found to increase with contact force [14]. That implies that excessive gate overdrive is not desired from a reliability standpoint. Recall from Figure 3.9 that the improvement of turn-on delay with gate overdrive saturates at ~$1.5V_{PI}$. Therefore, gate overdrive is kept to $1.5V_{PI}$ for all endurance measurements in this work to optimize endurance and speed.

Figures 3.11, 3.12. 3.13, and 3.14 show the evolution of R_{ON} with on/off cycling, comparing various conditions. The cycle-to-cycle fluctuations are caused by the formation and breakdown of native tungsten oxide. Figure 3.11 shows that V_B does not seem to affect R_{ON}. Although R_{ON} increases by ~1 order of magnitude over the course of 10^4 cycles, it is still within the acceptable range (< 100 kΩ). A comparison of hot and cold switching operations (Figure 3.12) show that cold switching operation results in faster increase and larger fluctuation in R_{ON} from one measurement to the next. As shown in Figure 3.13, lower drain bias (V_D = 1 V) leads to much faster increase in R_{ON} and larger fluctuations than larger drain bias (V_D = 2 V). In fact, R_{ON} shoots up from 1 kΩ to 10 MΩ in <50 cycles. These measurements confirm that native oxide continuously forms at the contact and needs to be broken down. When V_D = 2 V, we can see R_{ON} continuously rising and dropping again to lower values, indicating oxide formation and breakdown events. When V_D is too low, native oxide is broken further after the initial break, and reforms very quickly. It

should be noted that measurements conducted with high drain bias (V_D > 3V) increase the probability of welding induced failure, as drain current levels gets too high. It is therefore important to find the optimal drain voltage that would be able to minimize native oxide formation, but is still low enough to not cause welding over the device lifetime. V_D = 2 V is optimum for this contact technology. Figure 3.14 compares measurements conducted under different environments. No significant difference is observed between ambient and N_2 purge. Measurements conducted under vacuum (~10^{-6} Torr) shows much less R_{ON} fluctuation between measurements, which indicates much slower native oxide formation in low oxygen environment. On one device, cycling is stopped at various points to take I_D-V_G curves to investigate the evolution of V_{PI} and V_{RL} with cycling (Figure 3.15). The results show <1 V fluctuation in switching voltages with no particular trend up to 10^4 on/off cycles, so this design is structurally robust without dielectric charging concerns if gate overdrive is kept low.

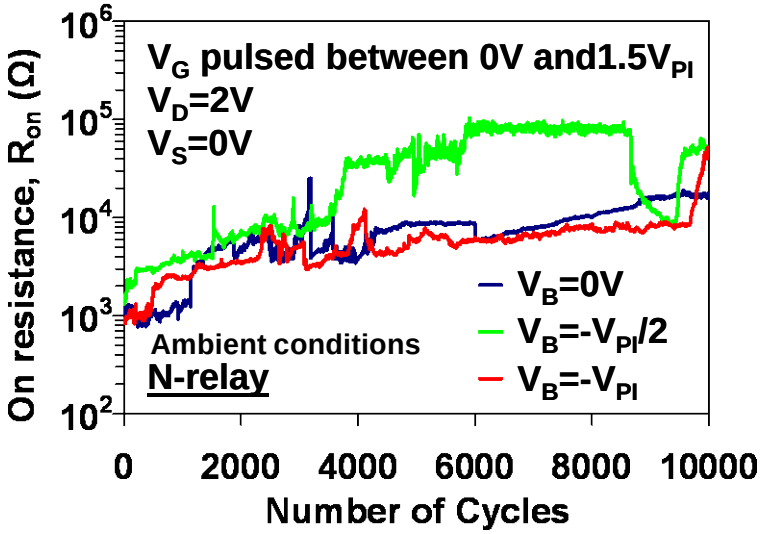


Figure 3.11: On-state resistance (R_{on}) with cycling for different applied V_B. For each measurement V_D = 2 V, V_S = 0 V were applied. V_G is pulsed from 0 V to 1.5V_{PI}. R_{on} for fresh devices are in the kΩ range. An increase in R_{on} to 10s of kΩ range is observed after 10,000 cycles. Statistical fluctuations in R_{on} observed with each cycle may be due to tungsten native oxide. Applied V_B does not appear to worsen the degradation of R_{on}.

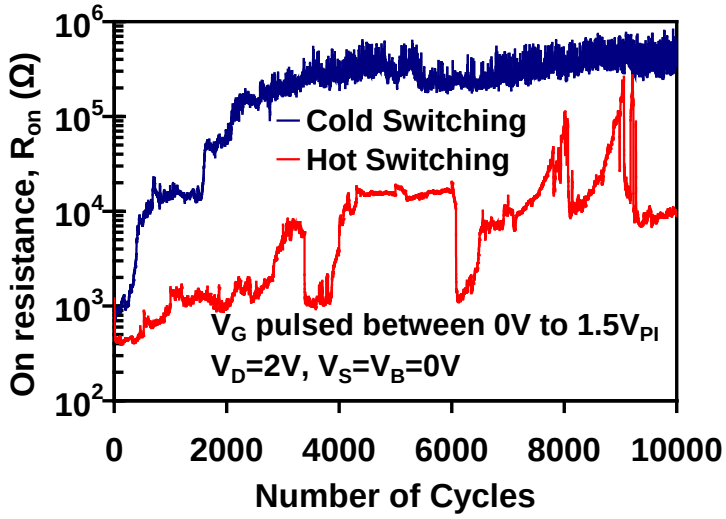


Figure 3.12: On-state resistance (R_{on}) with cycling for comparing hot and cold switching. In hot switching case, V_D = 2 V is applied at the time contact is made, followed by current measurement. For cold switching, V_D = 0 V at the time contact is made. V_D = 2 V is then applied for current measurement. For each measurement V_S = V_B = 0 V were applied. V_G is pulsed from 0 V to 1.5V_{PI}. R_{on} for fresh devices are in the kΩ range. An increase in R_{on} to 10s of kΩ range is observed after 10,000 cycles for hot switching. For cold switching, R_{on} increases much more rapidly to the 100s of kΩ range, indicating a more rapid tungsten oxidation at the contact.

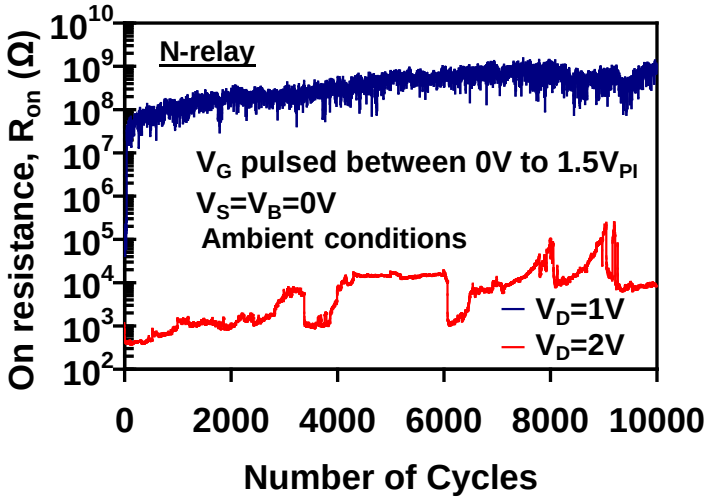


Figure 3.13: On-state resistance (R_{on}) with cycling for different V_D. For each measurement V_S = V_B = 0 V were applied. V_G is pulsed from 0V to 1.5V_{PI}. R_{on} for fresh devices are in the kΩ range. An increase in R_{on} to 10s of kΩ range is observed after 10,000 cycles for V_D = 2 V. R_{on} increases rapidly to 100s of MΩ range for V_D = 1 V, indicating oxidation of the tungsten native oxide at the contact.

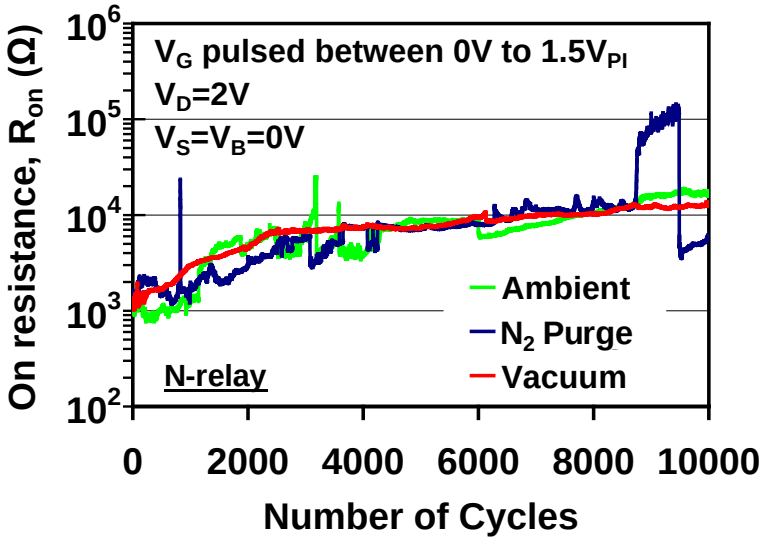


Figure 3.14: On-state resistance (R_{on}) with cycling tested under different environments. For each measurement V_D = 2 V, V_S = V_B = 0V were applied. V_G is pulsed from 0 V to 1.5V_{PI}. R_{on} for fresh devices are in the kΩ range. An increase in R_{on} to 10s of kΩ range is observed after 10,000 cycles. Statistical fluctuations in R_{on} observed with each cycle may be due to tungsten native oxide. Measurement under vacuum seems to help reduce this fluctuation in R_{on}, indicating slower rate of tungsten oxidation at the contact.

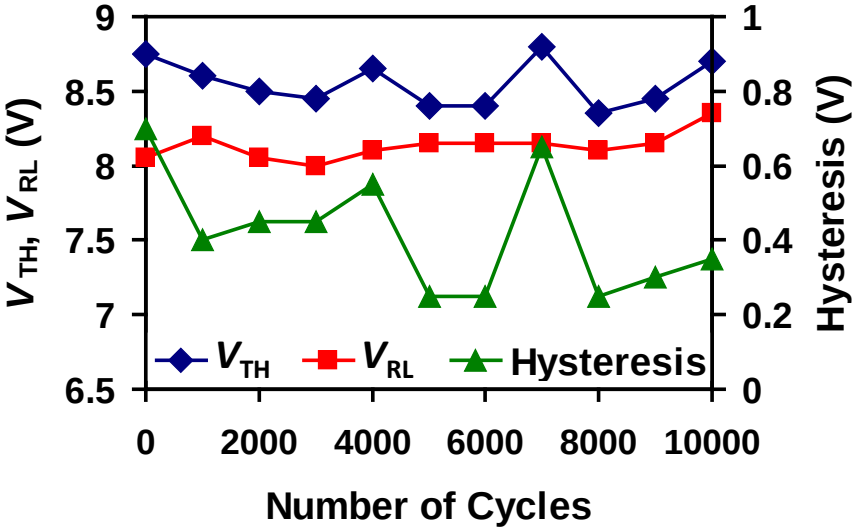


Figure 3.15: V_{PI}, V_{RL}, and hysteresis voltage (V_{PI}-V_{RL}) with cycling. V_G is pulsed from 0 V to 1.5V_{PI}, and V_D = 2 V, V_S = V_B = 0 V were applied during cycling. At regular intervals, cycling is stopped and I_D-V_G curve is measured to obtain V_{PI} and V_{RL}. Fluctuations in switching voltages are insignificant (<0.5V) without any trend for up to 10,000 on/off cycles.

To test for more cycles, measurements need to be conducted at higher frequencies. A simple way is to connect a function generator to the gate of the relay, and apply constant voltage to the drain, source and body. At specific intervals, the function generator is disconnected and I_{DS}-V_G curves are measured. Using this method, this relay design is shown to be able to switch >10^9 on/off cycles [4]. Atomic Force Microscopy (AFM) measurements of the contacting region (Figure 3.16) show that there is no measurable change in electrode topology (*i.e.* height profile remains unchanged) before and after 10^9 on/off cycles, affirming the superior wear properties of tungsten.

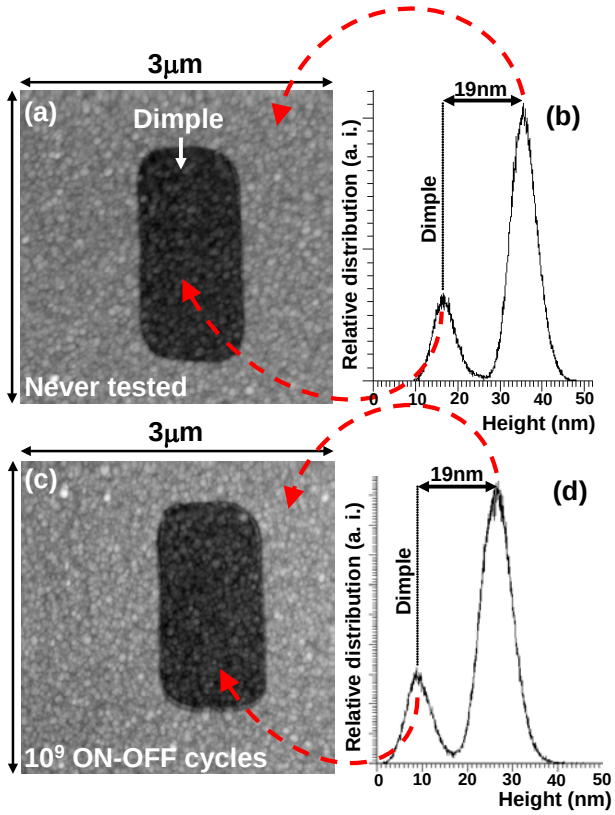


Figure 3.16: (a) AFM scan of a dimple contact (1μm x 2μm) region. The material is tungsten, coated with ALD TiO_2 (see Figure 3.3). (b) Relative distribution as a function of scan height. Measured dimple depth is 19 nm. This recessed (dark) region comes from process overetch into the underlying W when etching 1st sacrificial material to open the dimple regions. Figures (c) and (d) represent the same electrode structure, for a device cycled on/off 10^9 times at V_{DS} = 1 V and V_G =1.5V_{PI}. No wear is observed.

Reliability >10^{10} on/off cycles have also been achieved in subsequent tests [15]. Unfortunately, it is not practical to keep testing the relay to the point that it fails to switch. The maximum testing speed to conduct endurance measurements is limited by the switching speed of the relay. Based on turn-on delay measurements in Figures 3.9 and 3.10,

the maximum limit to testing frequency is ~1 MHz to guarantee that the relay turns on/off properly every cycle. Practically, the maximum number of cycles that can be experimentally measured is ~10^{11} to 10^{12} cycles (~1.6 to 11.6 days). In CMOS, accelerated lifetime testing is conducted to predict long term reliability typically by testing at voltages higher than the operating voltage and/or testing at higher temperatures. It does not work the same way for relays since the failure mechanisms are fundamentally different. A contact reliability model based on atomic diffusion was developed and experimentally verified to predict failure due to welding at the contact [16]. This model predicts reliability >10^{15} cycles for relays operating at 1 V.

An extensive study of R_{ON} evolution with cycling on W contacts show that this relay technology with W contacts can switch up to ~10^{8} on/off cycles under vacuum (5 μTorr) before the contact resistance shoots up above acceptable level for IC applications (10 kΩ for the study) [17]. Auger Electron Spectroscopy (AES) spectra measurements confirm that the increasing contact resistance is indeed caused by oxidation of the contacting surfaces.

3.4.4 Temperature & Radiation Effects

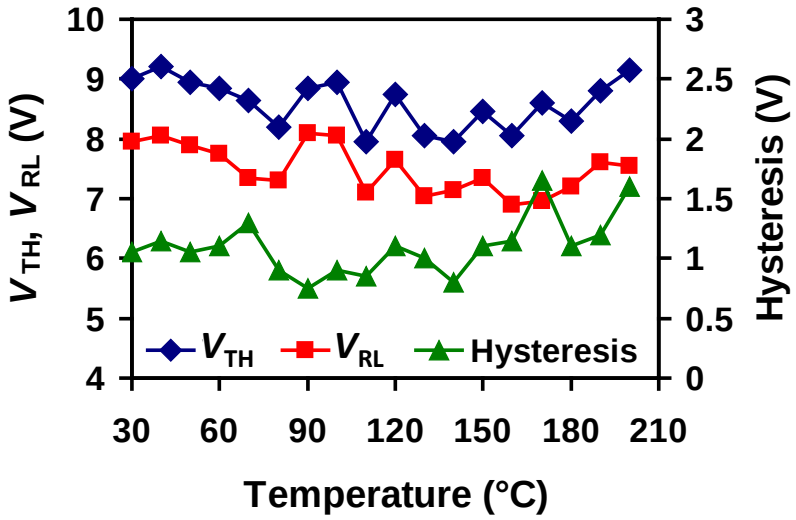


Figure 3.17: V_{PI}, V_{RL}, and hysteresis voltage (V_{PI}-V_{RL}) with temperature. Chuck temperature is increased from 30°C to 120°C with 10°C interval. At each interval, the chip is allowed to sit for 15 min before any measurement is made to allow temperature to stabilize. I_D-V_G curves is measured each interval with V_D = 2 V, V_S = V_B = 0 V to obtain V_{PI} and V_{RL}. Fluctuations in switching voltages are insignificant (<1 V) without any trend for the temperature range.

Another advantage micro-relays have over CMOS is robustness against temperature variations and radiation hardness, which is attractive especially for military and space applications. In orbit, thermal shocks can be severe (of order 16 cycles from -80°C to 100°C per day [18]). The effects of high temperature on V_{PI}, V_{RL}, and R_{ON} are investigated (Figures 3.17, 3.18). Within a temperature range of 30°-200°C, V_{PI} fluctuates between 8V-9.2 V, and R_{ON} between 1-15 kΩ, without a particular trend. This is still within the statistical fluctuation of the relay due to charging and contact oxidation. Low temperature measurements performed with the test setup cooled with liquid nitrogen show relays working as normal, but a systematic study was not performed. The folded flexure design is expected to be robust against temperature variation as it relieves thermal stress by expansion as shown by simulations in [4].

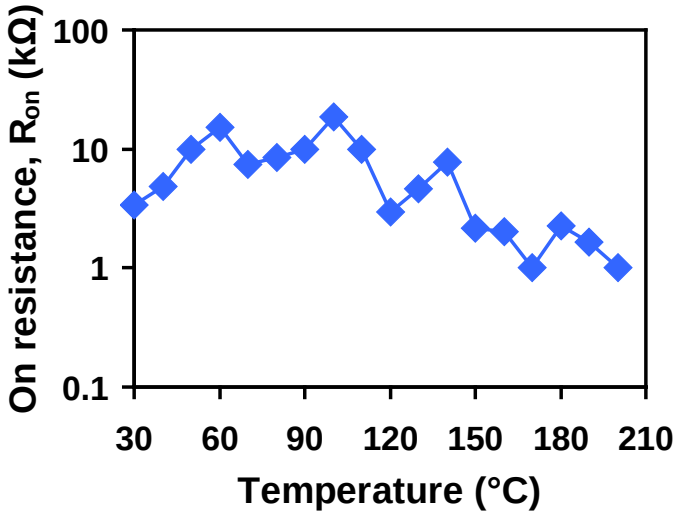


Figure 3.18: On-state resistance (R_{on}) of a 4T relay with temperatures. Chuck temperature is increased from 30°C to 120°C with 10°C interval. At each interval, the chip is allowed to sit for 15min before any measurement is made to allow temperature to stabilize. I_D-V_G curves is measured each interval with V_D = 2 V, V_S = V_B = 0 V to obtain R_{on}. R_{on} fluctuates between 1-15 kΩ range. High temperatures do not appear to worsen R_{on}.

MEMS devices are known to be able to withstand radiations from the 100s of krad to 10s of Mrad range [18]. To assess radiation hardness, the 4T relays are irradiated with alpha particles of various doses (200 krad, 2 Mrad, 20 Mrad). V_{PI} are measured before and after irradiation. Any shift in V_{PI} would imply charge accumulation at the 50nm Al_2O_3 gate dielectric. It is found that only relays exposed to 20 Mrad show ~10% jump in V_{PI}. Lower doses cause negligible shift in V_{PI}. Typical amount of radiation encountered in space due to cosmic radiation is ~50-100 rad/year [19], [20]. In the maximum intensity zone of inner Van Allen belt, radiation can get as high as 500 krad/year [20]. Even there, it

would still take 40 years to reach a level that shifts V_{PI} by 10% based on this estimate. Therefore, radiation is not expected to be a concern.

3.5 Parasitic Effects of 1st Generation 4T Relay Design

The design of the first working relay prototype is not perfect. Several undesirable phenomena were observed during device and circuit testing. Figure 3.19 shows the areas of the electrodes and the associated capacitances between them that exist in this relay design.

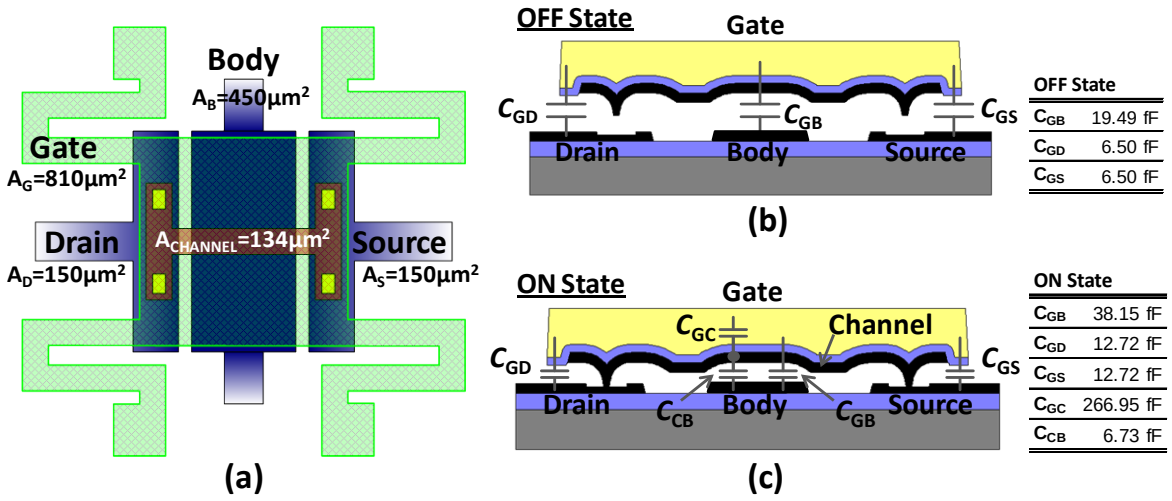


Figure 3.19: Capacitances associated with the 4T relay design. (a) Layout view, showing electrode areas. (b) Off-state and (c) on-state cross sectional views along the channel showing the associated capacitances. Their values are given in the table.

3.5.1 Actuation Asymmetry: Movable *vs.* Fixed Electrode

Throughout this chapter, the movable electrode (SiGe) has been biased as the gate. Since electrostatic force is ambipolar, the relay should also be able to be actuated by using the fixed electrode (W) equally effectively, as illustrated in Figure 3.20(a). In other words, using either the SiGe or W electrode as gate should result in the same V_{PI}. However, measured results shown in Figure 3.20(b) show that this is not the case.

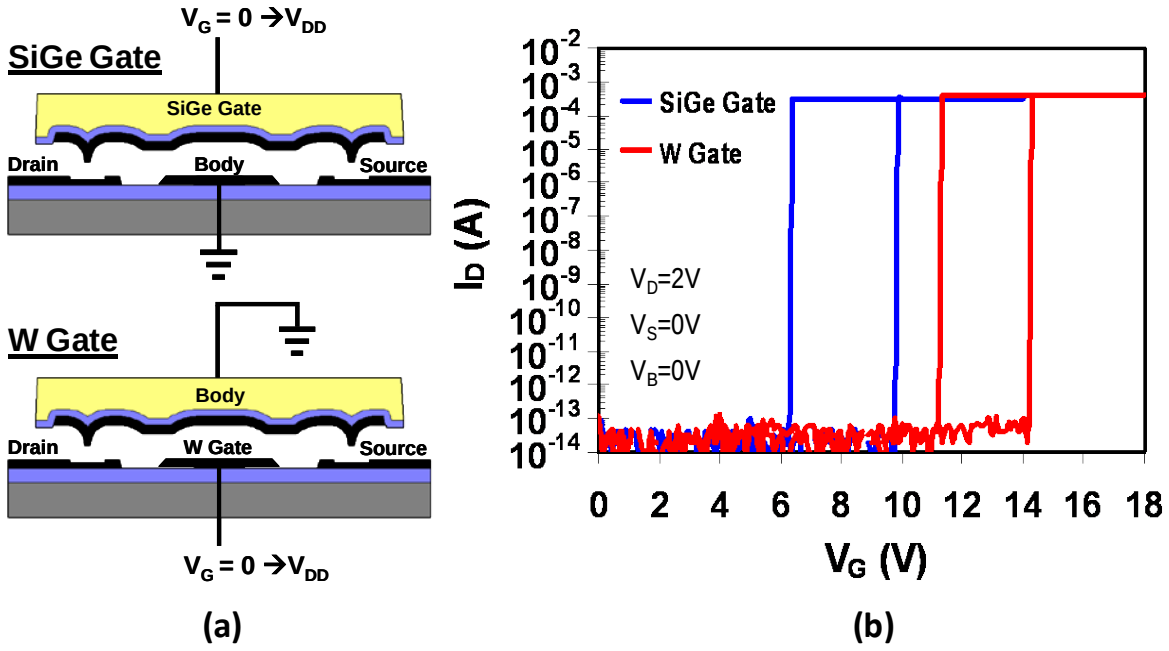


Figure 3.20: Since electrostatic force is ambipolar, either the movable (SiGe) electrode or the fixed (W) electrode can be biased as gate. (a) Bias configurations showing both bias cases. (b) I_D-V_G curves comparing a 4T relay biased with SiGe as gate *vs.* W as gate. For both measurements, V_D = 2 V and V_S = V_B= 0 V. V_{PI} and V_{RL} are significantly higher (by >4V) when W is biased as gate.

It is harder to actuate the relay using W as gate (V_{PI} ~ 14 V) *vs.* SiGe as gate (V_{PI} ~ 10 V). This can be attributed to the difference in size of the two electrodes. The area of the W electrode (A_W = 450 μm^2) is only about half the area of the SiGe electrode (A_{SiGe} = 810 μm^2). The electrostatic force (F_{elec}) between two plates is proportional to the capacitance and therefore the overlap area of the plates ($F_{elec} \propto C \propto A$). To a first order, all 810 μm^2 area of the movable plate contributes to electrostatic force when SiGe is biased to be the gate ($C_{ACTUATION} = C_{GB} + C_{GD} + C_{GS}$ = 32.5 fF). On the other hand, when W electrode is biased to be the gate, only 450 μm^2 total area contributes to electrostatic force ($C_{ACTUATION} = C_{GB}$ = 19.5 fF), resulting in a much weaker electrostatic force. While the current design necessitates the W electrode to always be smaller than the SiGe electrode to account for the source/drain regions, the area difference can be minimized in an optimally designed relay.

3.5.2 Body Effect

Figure 3.21 shows how the switching voltages shift with applied body bias. Ideally, actuation would depend only on the voltage difference between the gate and the body (V_{GB}), so V_{PI} and V_{RL} should shift the same amount as the applied V_B (*i.e.* slope = 1). However, everywhere a capacitance forms, electrostatic force exists to influence actuation. The slopes seen are instead ~0.5. Out of a total plate area of 810 μm^2, only about half (450 μm^2) forms capacitance with the body electrode (hence the slope of ~0.5). The gate-to-body electrode overlap should be maximized for body biasing to be more effective.

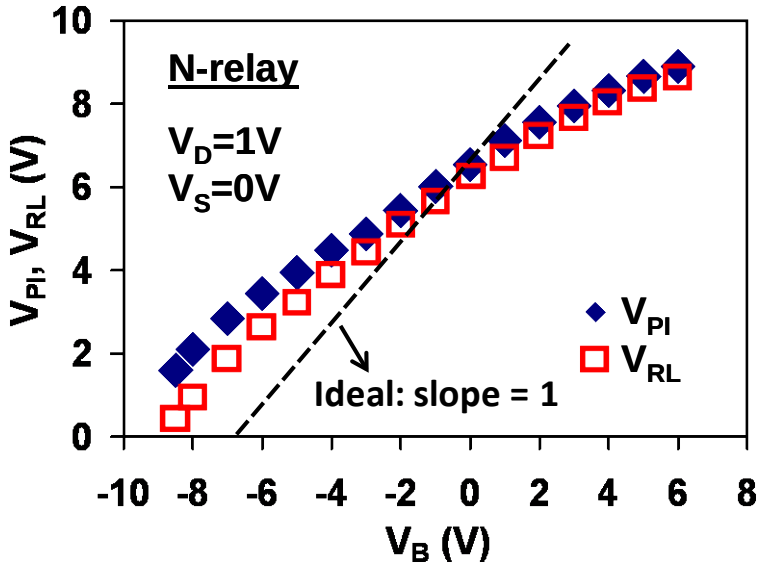


Figure 3.21: Dependence of the pull-in voltage (V_{PI}) and the release voltage (V_{RL}) on body bias (V_B). For a given V_G, more negative V_B results in larger V_{GB} (larger electrostatic force) and so reduces V_{PI} and V_{RL}, and vice versa. Slopes are 0.5046 and 0.5738 for V_{PI} and V_{RL} respectively. In the ideal case, the switching voltages should shift the same amount as the applied V_B (*i.e.* slope = 1).

3.5.3 Parasitic Source/Drain Actuation

Figure 3.22 shows that the drain and source bias voltages (V_D and V_S) also affect V_{PI} due to parasitic electrostatic forces. Ideally, the source and drain should have minimal effect to the switching voltages (*i.e.* slope = 0). The gate-to-source/drain overlap area (A_{GD} + A_{GS} = 300 μm^2) compared to gate-to-body overlap area (A_{GB} = 450 μm^2) is significant (~2/3). Thus, V_{PI} and V_{RL} are also influenced by V_D. This is a phenomenon similar to Drain Induced Barrier Lowering (DIBL) in CMOS, which causes V_{TH} lowering with higher

V_D. In an optimal relay design, the source and drain should not overlap with the actuated plate, so that A_{GD} and A_{GS} is minimized.

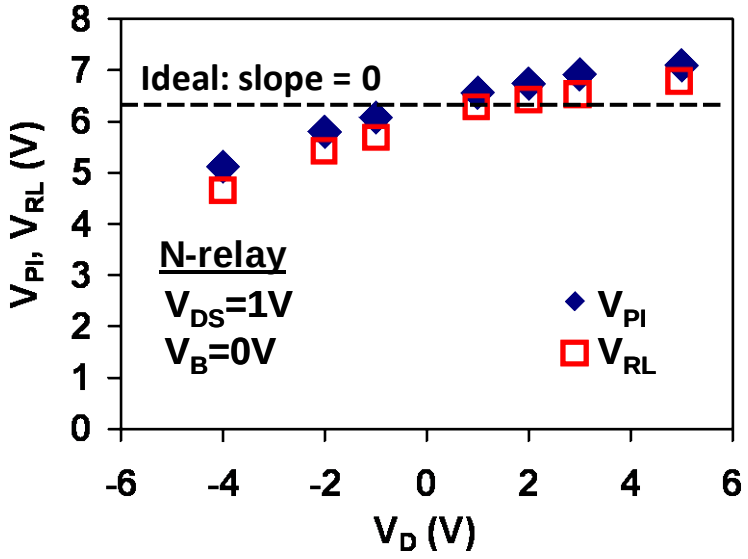


Figure 3.22: Dependence of the pull-in voltage (V_{PI}) and the release voltage (V_{RL}) on drain bias (V_D), with V_{DS} = 1V. Parasitic electrostatic force between the gate and the source/drain results in a shift in V_{PI} and V_{RL}. Slopes are 0.222 and 0.2377 for V_{PI} and V_{RL} respectively. In the ideal case, the switching voltages should not be affected by V_D and V_S (*i.e.* slope = 0).

3.5.4 Parasitic Channel Actuation

Another phenomenon that hampers circuit testing is the inability to turn off the relay in situations where both the source and drain are pulled high. Take a simple case of a relay-based buffer shown in Figure 3.23(a). When the input is high ($V_{IN} = V_{DD}$), the top relay will be on while the bottom relay will be off. The output is pulled up to high ($V_{OUT} = V_{DD}$). Therefore, both source and drain of the top relay is high (at V_{DD}). In cases such as this, the relay could not be turned off, even after V_G is lowered back to zero. As illustrated in 3.23(b), the channel potential is also at V_{DD} resulting in electrostatic force between the channel and the body. A large overlap area between the channel and the body (A_{CB}), creates electrostatic force large enough to overpower the spring restoring force (V_{RL} drops to below 0V). In cases where the channel voltage is not high enough to prevent turn off entirely, this phenomenon could still lower V_{RL} and undesirably increase the hysteresis voltage. Therefore, in a well designed relay, A_{CB} should be minimized, if not completely eliminated.

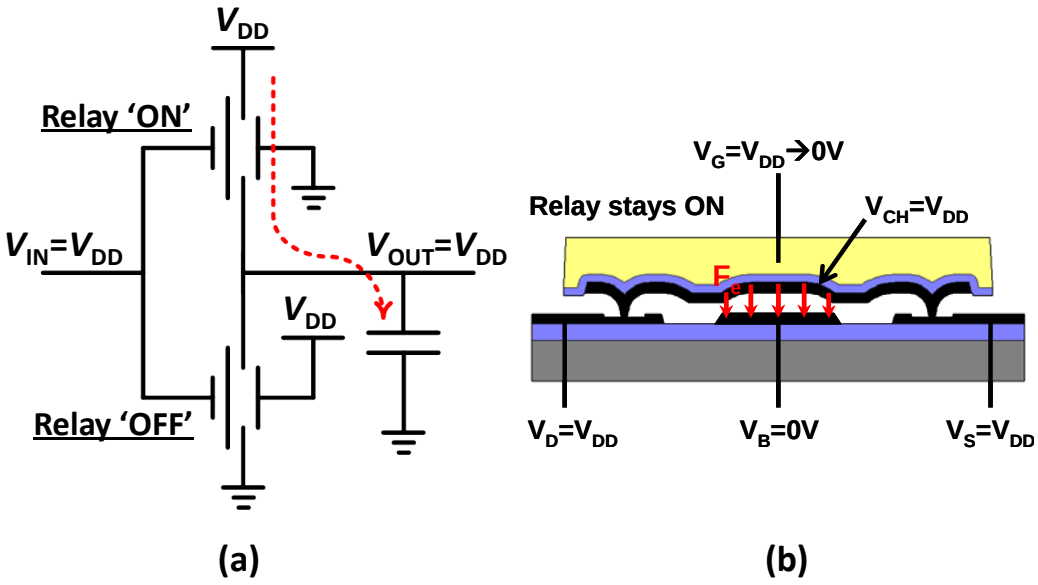


Figure 3.23: Description of parasitic channel actuation effect observed in the 1st generation 4T relay devices. (a) A buffer relay circuit, one example where this phenomenon can possibly happen. (b) Bias configuration that prevents the device from turning off despite V_G lowered back to 0V. This is caused by significant electrostatic force due to large channel-to-body overlap area.

3.6 References

[1] K. Akarvardar, D. Elata, R. Parsa, G. C. Wan, K. Yoo, J. Provine, P. Peumans, R. T. Howe, and H.-S. P. Wong, "Design considerations for complementary nanoelectromechanical logic gates," in Proc. International Electron Devices Meeting, pp. 299-302, 2007.

[2] F. Chen, H. Kam, D. Markovic, T. King Liu, V. Stojanovic, E. Alon, "Integrated Circuit Design with NEM Relays." IEEE/ACM International Conference on Computer-Aided Design, pp 750-757, Nov. 2008.

[3] H. Kam, T.-J. King Liu, E. Alon, and M. Horowitz, "Circuit level requirements for MOSFET-replacement devices," in Proc. International Electron Devices Meeting, pp.427, 2008.

[4] H. Kam, V. Pott, R. Nathanael, J. Jeon, E. Alon, and T.-J. King Liu, "Design and reliability of a micro-relay technology for zero-standby-power digital logic applications," in Proc. International Electron Devices Meeting, pp. 809-812, 2009.

[5] V. Pott, H. Kam, J. Jeon, and T.-J. King Liu, “Improvement in mechanical contact reliability with ALD TiO2 coating,” in Proc. AVS Conference, pp. 208-209, 2009.
[6] K. R. Williams and R. S. Muller, “Etch rates for micromachining processing,” Journal of Microelectromechanical Systems, vol. 5, no. 4, pp. 256-269, 1996.
[7] R. Holm, Electric Contacts: Theory and Applications, Berlin, NY: Springer-Verlag, 1967.
[8] C. W. Low, T.-J. King Liu, and R. T. Howe, “Characterization of polycrystalline silicon-germanium film deposition for modularly integrated MEMS applications,” Journal of Microelectromechanical Systems, vol. 16, no. 1, pp. 68-77, Feb. 2007.
[9] G. M. Rebeiz, “RF MEMS: Theory, Design, and Technology,” New York: John Wiley & Sons, 2003.
[10] J. Jeon, W. Kwon, and T.-J. K. Liu, "Embedded memory capability of four-terminal relay technology," IEEE Transactions on Electron Devices, vol. 58, no. 3, pp. 891-894, 2011.
[11] G. Gielen, P. De Wit, E. Maricau, J. Loeckx, J. Martín-Martínez, B. Kaczer, G. Groeseneken, R. Rodríguez, and M. Nafría, “Emerging yield and reliability challenges in nanometer CMOS technologies,” in Proc. IEEE Design, Automation and Test in Europe, pp. 1322–1327, Mar. 2008.
[12] W. Merlijn van Spengen, “MEMS reliability from a failure mechanisms perspective." Microelectronics Reliability,” vol. 43, no. 7, pp. 1049-1060, 2003.
[13] R. Modlinski, A. Witvrouw, A. Verbist, R. Puers, and I. De Wolf, “Mechanical characterization of poly-SiGe layers for CMOS–MEMS integrated application,” Journal of Micromechanics and Microengineering, vol. 20, no. 1, 2009.
[14] D. Lee, V. Pott, H. Kam, R. Nathanael, T.-J. K. Liu, "AFM characterization of adhesion force in micro-relays," 2010 IEEE 23rd International Conference on Micro Electro Mechanical Systems, pp. 232-235, 2010.
[15] Y. Chen, University of California, Berkeley, unpublished work.
[16] H. Kam, E. Alon, and T.-J. K. Liu, "A predictive contact reliability model for MEM logic switches," IEEE International Electron Devices Meeting Technical Digest, pp. 399-402, 2010.
[17] Y. Chen, R. Nathanael, J. Jeon, J. Yaung, L. Hutin, and T.-J. K. Liu, “Characterization of contact resistance stability in MEM relays with tungsten electrodes,” IEEE/ASME Journal of Microelectromechanical Systems, vol. 21, no. 3, pp. 511-513, 2012.
[18] H. R. Shea, “Reliability of MEMS for space applications,” in Proc. SPIE, vol. 6111, p. 61110A, 2006.
[19] R. A. Mewaldt, A. J. Davis, W. R. Binns, G. A. de Nolfo, J. S. George, M. H. Israel, R. A. Leske, E. C. Stone, M. E. Wiedenbeck, and T. T. von Rosenvinge, “The Cosmic Ray Radiation Dose in Interplanetary Space Present Day and Worst-Case Evaluations,” in International Cosmic Ray Conference, vol. 2, p. 101-104. 2005.
[20] T. Foelsche, “Estimates of radiation doses in space on the basis of current data,” Life sciences and space research 1, pp. 48-94, 1963.

Chapter 4

Relay Process & Design Optimization for Low Voltage Operation

4.1 Introduction

The ideal switching characteristic of zero off-state leakage and abrupt switching behavior of relays makes it attractive for digital integrated circuit applications as chip power density has become a major challenge in recent years. A robust and reliable 4-Terminal (4T) relay technology, having high yield (>95%) and excellent endurance (>10^{10} on/off cycles) has been demonstrated in the previous chapter of this dissertation. Indeed, ideal switching characteristic is demonstrated with SS < 0.1mV/dec and immeasurably low I_{OFF} (<10^{-14}A). However, the process technology and design are not optimized. The device footprint is unnecessarily large, and the release process relies on a timed etch. Parasitic electrostatic effects were found and operating voltages remain too large, making it difficult to implement complex logic circuits with the technology. In order to fully realize the promise of relays as an alternative to CMOS for low power digital circuits, the relays need to be able to operate with voltages <1 V, *i.e.* with minimal hysteresis. The hysteresis voltage sets the minimum V_{PI} where the relays can still turn off (*i.e.* when V_{RL}=0V, V_{PI} = hysteresis voltage).

This chapter focuses on relay process and design optimization. With optional additional lithography steps, we can achieve a more robust process required for eventual device scaling, highly reduced device footprint area, and the ability to form interconnects to fabricate circuits. Improved relay designs minimize parasitic electrostatic force issues and enhance functionality. The process flow is optimized to achieve the lowest switching voltages, while still ensuring reliable turn-off and avoiding stiction. New device concepts increase device functionality to reduce the number of structures required to achieve a certain function, and provide a new paradigm for designing circuits using relays. Test chips were fabricated containing arrays of devices of various designs to find the optimal design trading off gate control, switching voltages, and reliability.

4.2 5-Mask/7-Mask Process

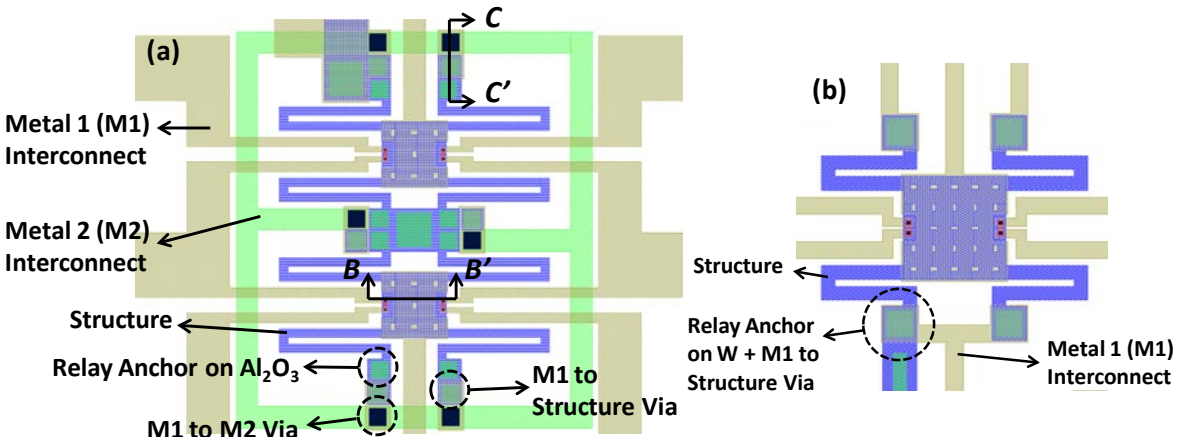
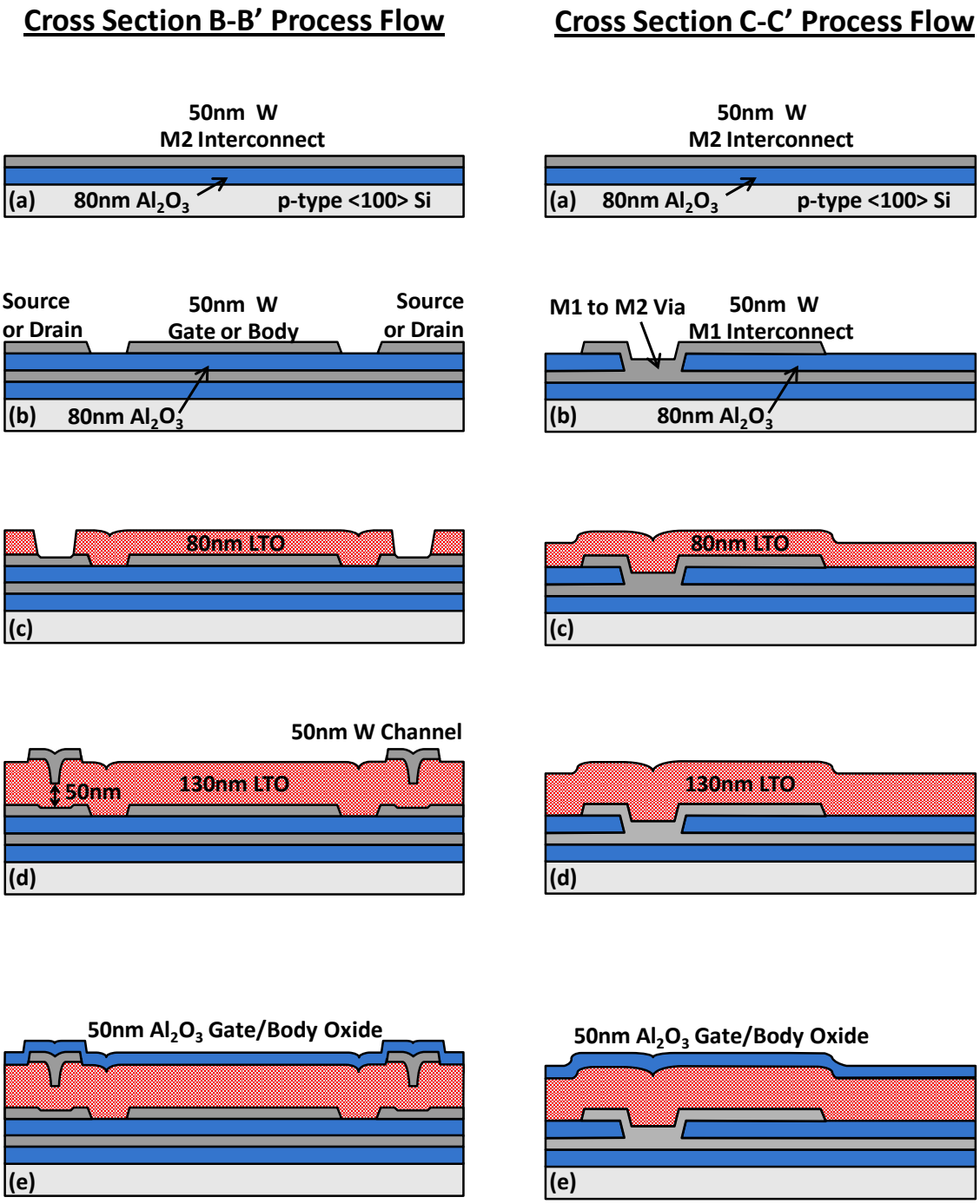


Figure 4.1: Layout view illustrating improved anchor and interconnect design on dual-source/drain devices. **(a)** An inverter circuit. Relays are anchored on Al_2O_3 substrate dielectric. M1 to structure via enables connection from the SiGe structure to the W fixed electrodes. M1 to M2 via provides a connection to an interconnect layer underneath the relay. All flexures are connected through the M2 interconnect, a low resistance path to minimize voltage drops throughout the structure. **(b)** Device anchored directly on W (M1 interconnect). Combining anchor and via saves area.

As we attempt to design for lower voltages, an improvement to the process is required. The 4-mask process shown in Chapter 3 is simple and sufficient to fabricate the first prototype devices. However, it can be improved to facilitate miniaturization and to fabricate large circuits reliably. In the original process, the structural layer is anchored on the sacrificial material. As a result, the release time needs to be kept short and the printed anchor size needs to be large to make sure the anchors survive. With scaling, gap thicknesses are expected to be smaller and a long release time may be required. A timed release requirement could potentially limit scaling. Large anchor regions also consume a lot of area, unnecessarily increasing the device footprint. In fact, the anchors (50 μm × 50 μm each) make up >50% of the total device area (120 μm × 150 μm) in the 1st generation design. The location and size of the actual anchor is not lithographically fixed. It is determined by the lateral etch distance of the sacrificial oxide underneath, potentially an added source of process-induced variation in the effective spring constant. An extra mask (*i.e.* a 5-mask process) is needed to form a more reliable anchor and to electrically connect the structural layer and the fixed electrodes to build circuits. The new anchor and interconnect design are illustrated in Figure 4.1. Anchors can be formed directly on the Al_2O_3 substrate dielectric or the W electrodes (M1 layer) to form an electrical connection for circuit routing. The anchor size is lithographically defined, the HF vapor etch time is

not limited, and large anchors are not required. An anchor size of 10 μm × 10 μm is adequate, reducing the device footprint to 68 μm × 78 μm. An optional two additional masks (*i.e.* a 7-mask process) could be used to provide an additional layer of interconnect (M2 layer) for larger circuits.



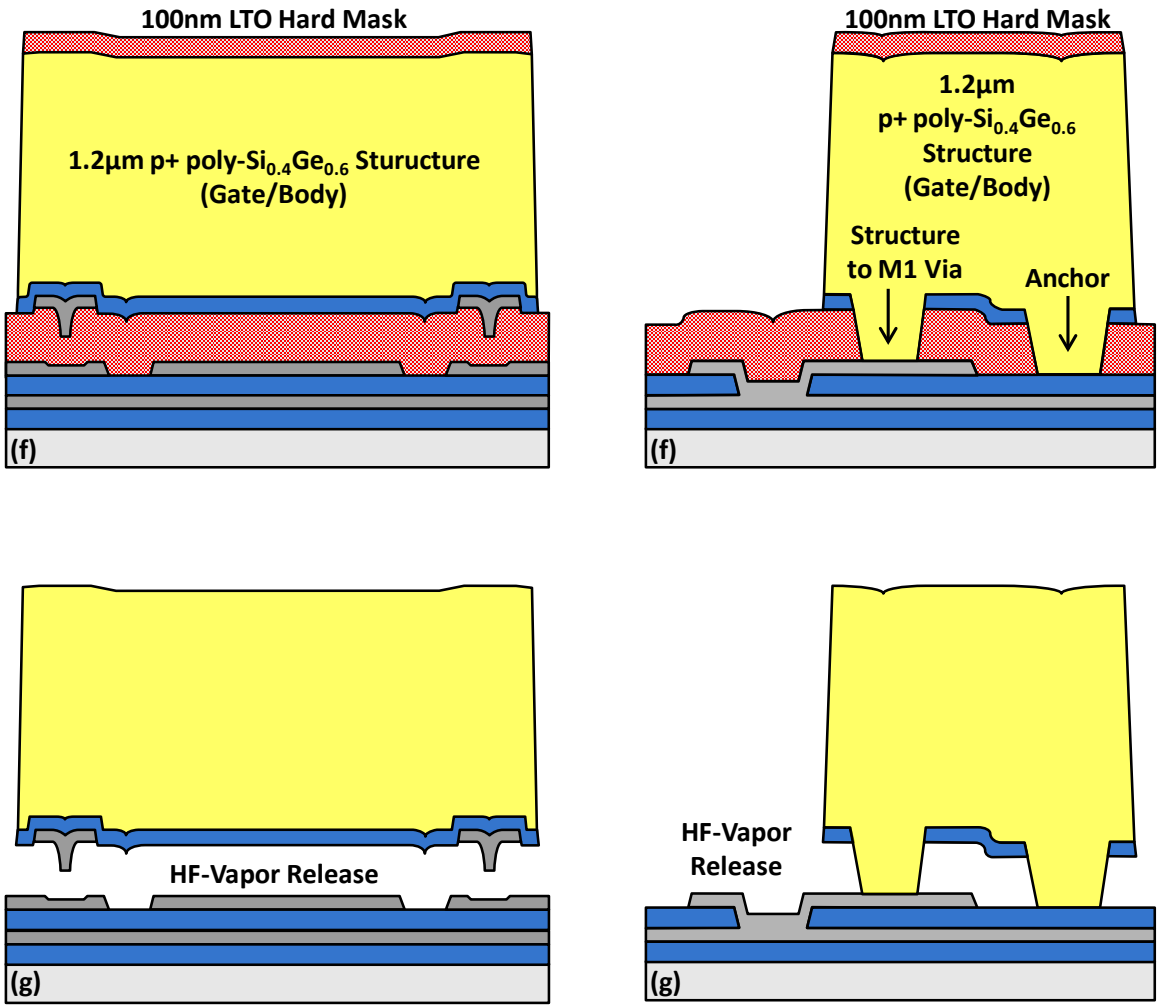


Figure 4.2: Illustrations along cross section B-B' and C-C' in Figure 4.1, showing a 5-mask/7-mask process flow for fabricating a four-terminal relay: **(a)** Underlying tungsten interconnect layer (M2) is deposited on Al_2O_3-coated substrate. **(b)** M2 layer is covered by Al_2O_3 inter layer dielectric (ILD), vias are etched, and tungsten fixed electrodes (gate/body, source, and drain) electrodes are formed. **(c)** Deposition of 1st sacrificial Low-Temperature-Oxide (LTO) layer and definition of source/drain contact regions. **(d)** Deposition of 2nd LTO layer and formation of W channels. **(e)** Deposition of Al_2O_3 body oxide. **(f)** Formation of relay anchor and structure to M1 via, followed by deposition of p+ poly-$Si_{0.4}Ge_{0.6}$ body layer and LTO hard mask, and patterning of body electrode and body oxide. **(g)** HF-vapor release.

The improved process is illustrated in Figure 4.2. On a starting silicon wafer substrate, a layer of Al_2O_3 (80 nm) is deposited by ALD to form an insulating substrate surface. A 50 nm tungsten layer is deposited by sputtering and patterned to form Metal 2 (M2) interconnect layer. Another 80 nm layer of Al_2O_3 is deposited by ALD as the inter-layer dielectric (ILD) and M1-to-M2 via holes are opened. A second tungsten layer is deposited by sputtering and patterned to form Metal 1 (M1) interconnect, the gate/body, source, and drain electrodes. Next, a first sacrificial layer of SiO_2 is deposited via LPCVD and patterned to define the contacting regions. Due to overetching, the surface of the tungsten source/drain electrodes is slightly recessed. A second sacrificial layer of SiO_2 is deposited. The 2nd sacrificial layer thickness defines the contact dimple gap, whereas the total thickness of the two sacrificial layers defines the actuation gap. Next, a tungsten layer is deposited and patterned to form the channel. Afterwards, a layer of Al_2O_3 is deposited to form the insulating gate/body oxide layer, followed by definition of structure-to-M1 via holes. P^+ poly-$Si_{0.4}Ge_{0.6}$ structural layer is deposited by LPCVD and patterned along with the gate/body oxide layer with the aid of SiO_2 hard-mask layer. The structures are released by selectively etching away all of the SiO_2 in HF vapor. As an optional step, an ultra-thin layer of TiO_2 can be conformally deposited by ALD for reliability improvement.

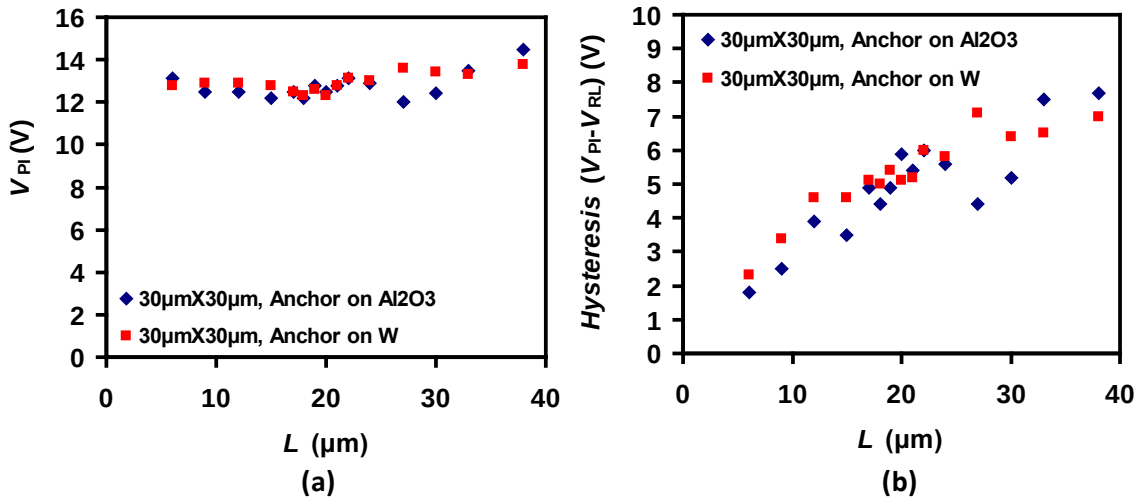


Figure 4.3: Comparison of a 4T relay design anchored on Al_2O_3 and W showing **(a)** V_{PI} and **(b)** hysteresis voltage. No difference is seen. This affirms that the relay structure can be anchored directly on M1 (*i.e.* the anchor and via can be combined) to reduce circuit layout area.

Note that in this chapter, the gate and body electrodes are interchangeable. It is beneficial for circuit design to be able to use either the movable (SiGe) electrode or fixed (W) electrode to actuate the relay. Measurements are performed to see if anchoring on

Al_2O_3 or W would make a difference for a range of different flexure lengths (L) (Figure 4.3). Very similar V_{PI} and hysteresis are measured. Thus, the relay anchor can also be used simultaneously as a structure-to-M1 connection to save area.

4.3 Relay Designs for Improved Electrostatics

4.3.1 2nd Generation 4-Terminal Relay Design

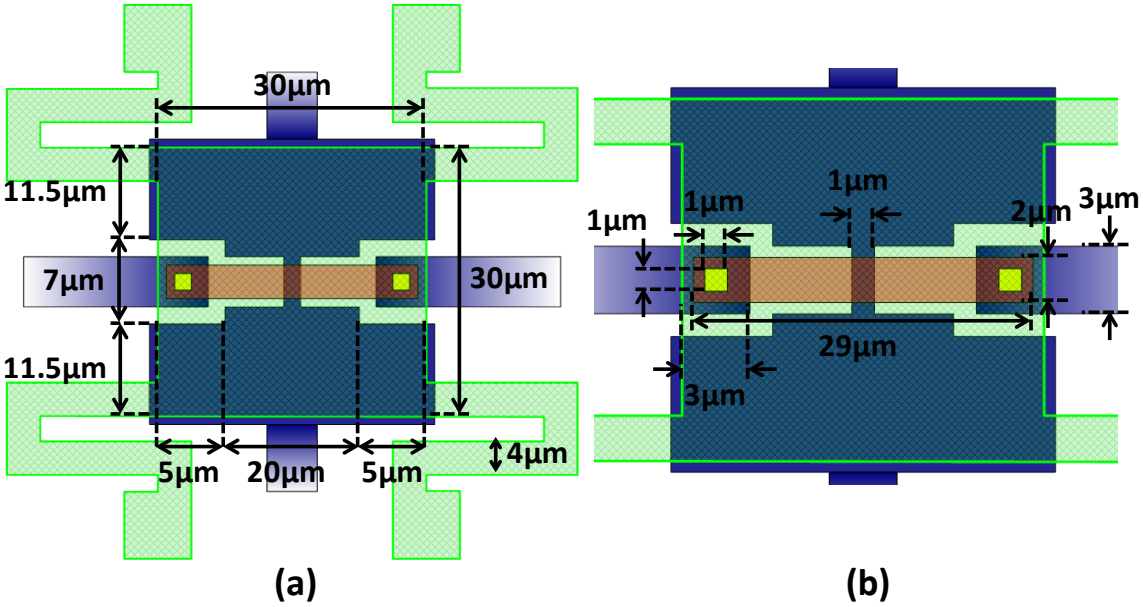


Figure 4.4: 2nd generation 4T relay design. **(a)** Key dimensions. **(b)** Zoomed-in view showing channel and dimple regions. Total footprint size is 68µm×78µm.

Having identified several issues in the 1st generation design (Section 3.5), the relay design is improved with the following in mind:

(1) Maximize gate-to-body overlap area
(2) Minimize gate-to-source/drain overlap area
(3) Minimize channel-to-body overlap area

The 2nd generation is shown in Figure 4.4. The key is to minimize the source/drain area that penetrates underneath the actuation area. The limit to keep reducing the source/drain area is the dimple size, which is lithographically limited. The ASML 5500/300 deep ultraviolet (DUV) stepper used to fabricate these relays has a minimum resolution of 250 nm if fully optimized. However, it is harder to etch small holes than to pattern small features. A 1 μm × 1 μm dimple size is very robust against process variations. With careful optimization (focus/exposure, exposure energy), smaller dimples can be made. In this work, dimple sizes down to 0.4 μm × 0.4 μm are made with reasonably good yield. The fixed electrode under the channel is removed to minimize channel-to-body overlap area, leaving a small (minimum-size) strip to electrically connect the top and bottom halves of the fixed electrode.

As shown in Figure 4.5, the switching voltages differ by <1 V when using either the SiGe electrode or W electrode as gate (*vs.* ~4-5V in 1st generation design). With SiGe as gate, Figure 4.6 shows that the slopes of V_{PI} and V_{RL} *vs.* V_B are close to ideal (~1). Body biasing now shifts the switching voltages more effectively. Parasitic actuation from the source/drain regions are also eliminated as the slopes of V_{PI} and V_{RL} *vs.* V_D is ~0, shown in Figure 4.7. The inability to turn the relays off due to parasitic channel actuation is no longer observed during subsequent device and circuit testing. Simple logic, clocking and memory circuits have been demonstrated with the 1st generation design [1]. With improved electrostatic control, the 2nd generation design enables demonstrations of more complex circuits, such as adder [2] and multiplier [3], and also power gating [4]. The largest circuit demonstrated to date consists of 98 working relays in a circuit [3].

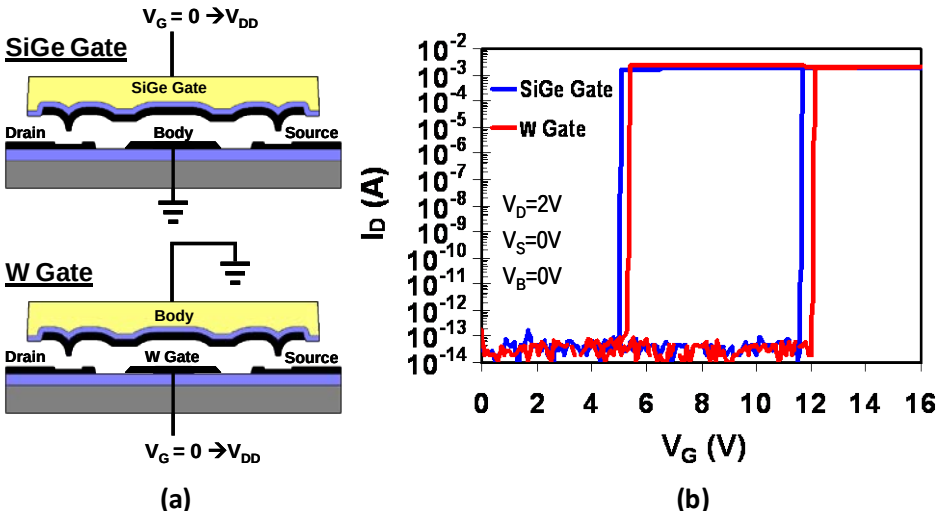


Figure 4.5: Since electrostatic force is ambipolar, either the movable (SiGe) electrode or the fixed (W) electrode can be biased as gate. **(a)** Bias configurations showing both bias cases. **(b)** I_D-V_G curves comparing a 2nd generation 4T relay biased with SiGe as gate *vs.* W as gate. For both measurements, V_D = 2 V and V_S = V_B = 0 V. V_{PI} and V_{RL} for both cases differs by <0.5V, significant improvement from the 1st generation design.

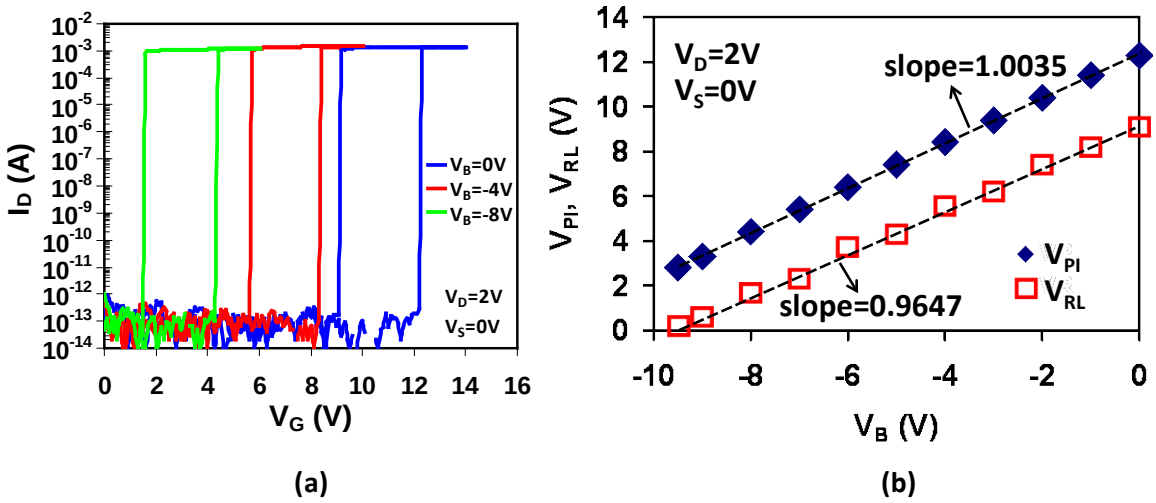


Figure 4.6: **(a)** Measured I_{DS}-V_G characteristics of a 2nd generation 4T relay with different body biasing. V_D = 2 V, V_S = 0 V, and V_B = 4 V, 0 V, -4 V, and -8 V. **(b)** Dependence of the pull-in voltage (V_{PI}) and the release voltage (V_{RL}) on body bias (V_B). For a given V_G, more negative V_B results in larger V_{GB} (larger electrostatic force) and so reduces V_{PI} and V_{RL}, and vice versa. Slopes are 1.0035 and 0.9647 for V_{PI} and V_{RL} respectively, which is close to the ideal case (*i.e.* slope = 1). Gate is the movable (SiGe) electrode.

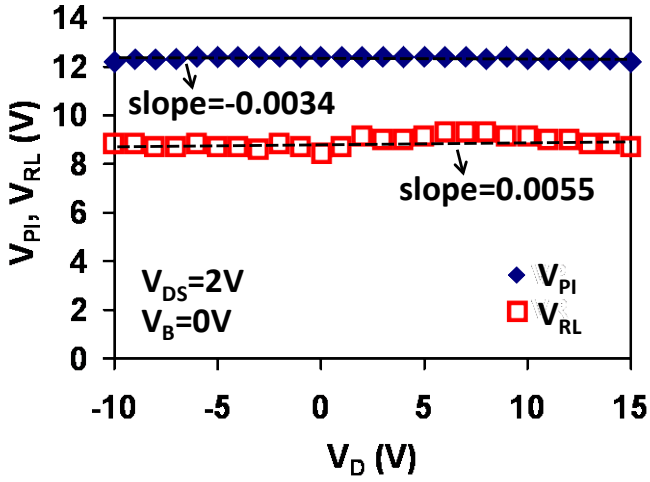


Figure 4.7: Dependence of the pull-in voltage (V_{PI}) and the release voltage (V_{RL}) on drain bias (V_D), with V_{DS} = 2 V. Parasitic electrostatic force between the gate and the source/drain results in a shift in V_{PI} and V_{RL}. Slopes are -0.0034 and 0.0055 for V_{PI} and V_{RL} respectively, which is close to ideal (*i.e.* slope = 0). The 2nd generation design has eliminated source/drain parasitic actuation effect. Gate is the movable (SiGe) electrode.

Note that the pull-in voltages are higher in this technology compared to the 1st generation design. Previously, parasitic actuation from the source/drain regions helped to lower V_{PI}. An appreciable amount of body area is also lost on the 2nd generation design to minimize channel-to-body overlap. While we have gained better control over the operation of the device, V_{PI} is still too high for reliable circuit operation and the ultimate goal of low power operation. Process optimizations are attempted to reduce V_{PI}. Recall from Chapter 1, V_{PI} depends on:

$$V_{PI} \propto \sqrt{\frac{k_{eff} {g_0}^3}{A_{eff}}} \qquad \text{when } g_d \geq \frac{1}{3} g_0 \qquad (4.3.1.1)$$

$$V_{PI} \propto \sqrt{\frac{k_{eff} g_d (g_0 - g_d)^2}{A_{eff}}} \qquad \text{when } g_d < \frac{1}{3} g_0 \qquad (4.3.1.2)$$

g_0 and g_d are actuation and dimple gaps respectively. A_{eff} can be approximated as the overlap area between the movable and fixed electrode. For this particular relay design, the effective spring constant (k_{eff}) consists of flexural and torsional components, and can be approximated as follows [5]:

$$\frac{1}{k_{eff}} \cong \left(\gamma_f \frac{EWh^3}{L^3}\right)^{-1} + \left(\gamma_t \frac{GWh^3}{L}\right)^{-1} \qquad (4.3.1.3)$$

where γ_f is the flexural constant, γ_f is the torsional constant, E is the Young's modulus, G is the shear modulus, and L, W, and h are the length, width, and height of the flexures respectively. While W and L are defined by layout, h is the structural SiGe thickness (T_{SiGe}) defined by process. Several relay lots were fabricated with various splits to optimize the process. The gap sizes (g_0 and g_d) and structural layer thickness (T_{SiGe}) are varied. A total of 25 wafers were completed and the results are summarized in Figure 4.8.

Devices with 30nm 2nd sacrificial layer, thus thin contact gap (g_d = 30 nm) (Wafers #2-4) are always in the on-state (short between source and drain even when V_G = 0 V). These relays are stuck down during release as spring restoring force is not enough to overcome surface forces. The smallest contact gap that can be fabricated successfully is 50 nm.

Devices with 50 nm 1st sacrificial layer (Wafers #13-16) do not show current conduction. The 1st sacrificial layer determines the height of the contact dimple. A dimple height of 50 nm may not be enough to prevent the gate dielectric "foot" problem (Section 2.4.4), due to process-induced variations.

	Process Parameters (nm)					Results	
Wafer #	T_{SiGe}	1st sac	2nd sac	g_0	g_d	V_{PI}	Note
1	1000	100	100	200	100	12-13V	
2	1000	90	30	120	30	-	Short
3	1000	110	30	140	30	-	Short
4	1000	150	30	180	30	-	Short
5	900	110	120	230	120	13-14V	
6	900	110	90	200	90	12-13V	
7	900	100	80	180	80	12V	
8	900	90	90	180	90	12V	
9	900	75	65	140	65	12V	
10	900	75	85	160	85	12-13V	
11	900	75	95	170	95	12-13V	
12	900	75	110	185	110	13-14V	
13	900	50	30	80	30	-	Open
14	900	50	50	100	50	-	Open
15	900	50	80	130	80	-	Open
16	900	50	100	150	100	-	Open
17	550	110	110	220	110	20-23V	
18	750	110	110	220	110	14-17V	
19	1000	110	110	220	110	12-13V	
20	1000	75	75	150	75	11-12V	
21	1200	75	75	150	75	11V	
22	1200	75	85	160	85	11-12V	
23	1200	80	50	130	50	10V	
24	1200	80	80	160	80	10V	
25	1200	80	100	180	100	11-12V	

Figure 4.8: Process optimization using 2nd generation 4T relay design to achieve low voltage operation. Structural SiGe thickness (T_{SiGe}) and sacrificial layer thicknesses, which determine actuation gap thickness (g_0) and contact gap thickness (g_d), are the process parameters of interest. 25 wafers are fabricated and characterized. V_{PI} corresponds to typical values taken from relays with flexure lengths between 15 μm to 22 μm, taking into account variations. V_{PI} is measured with the movable electrode (SiGe) as gate.

While gap thicknesses affect V_{PI}, their effect seems to be kept to a small (~2 V) range for a given T_{SiGe}. For example, for T_{SiGe} = 900 nm, V_{PI} is always between 12 V and 14 V. No matter how much we reduce the actuation gap thickness, V_{PI} could not be reduced under 12 V. On the other hand, T_{SiGe} appears to have a big impact on V_{PI}. From 550 nm to 1200 nm, V_{PI} dramatically reduces with thicker T_{SiGe}. This is counterintuitive since thicker SiGe would make a stiffer beam. However, thinner films show dramatic

increase in strain gradient whose effect dominates that of the effective spring constant. Very large negative strain gradient is observed for T_{SIGE} < 900 nm, leading to very large V_{PI}. Warping is clearly visible under microscope, with the actuation plate lifted upwards. A rapid thermal anneal (RTA) step is done for 1 minute at 600°C (maximum allowable temperature for tungsten) for devices with thin SiGe. Strain gradient is found to worsen with this RTA step. V_{PI} shoots up to 30-35V and 16-18V for Wafers #17 and #18 respectively.

The optimal structural SiGe thickness is 1200 nm, resulting in V_{PI} between 10V and 12V. Thus far, V_{PI} = 10 V is the lowest switching voltage achieved with this relay process (Wafers #23-24). Interferometer measurements (Figure 4.9) show that the plate and flexure are relatively flat with sufficiently low strain gradient extracted to be ~1.7×10^{-4}/µm.

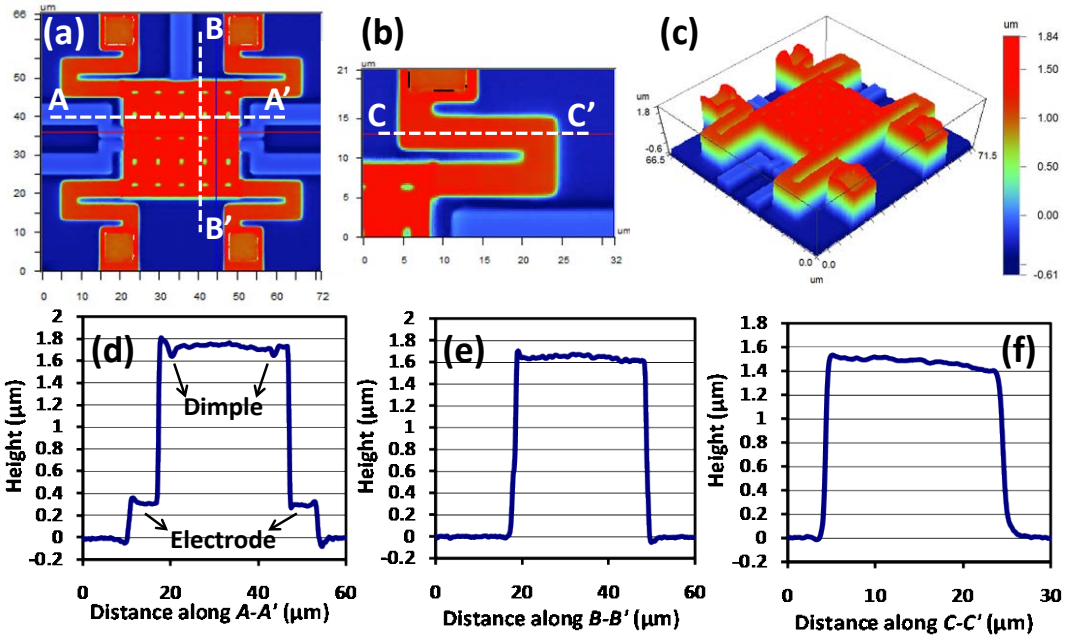


Figure 4.9: Interferometry analyses of a relay with 1200 nm-thick poly-SiGe, after HF-vapor release: **(a)** plan view, **(b)** zoomed-in view of a flexure and **(c)** 3-dimensional view. **(d)** Height profiles along *D-D'*, **(e)** *E-E'* and **(f)** *F-F'*. Extracted strain gradient ~1.7×10^{-4}/µm.

Strain gradient presents a significant challenge for fabricating devices that operate at low voltages. Not only does it restrict V_{PI} to a certain range, it also makes V_{PI} difficult to predict since the models are no longer accurate. This highlights the importance of developing a material with minimal strain gradient.

4.3.2 3rd Generation 4-Terminal Relay Design

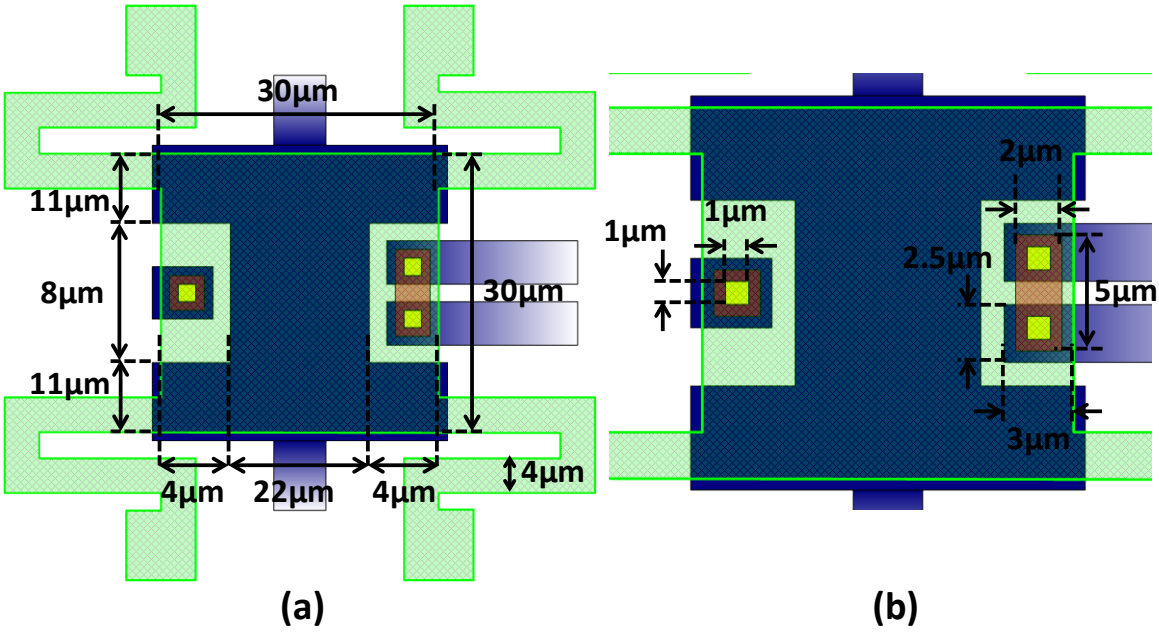


Figure 4.10: 3rd generation 4T relay design. **(a)** Key dimensions. **(b)** Zoomed-in view showing channel and dimple regions. Total footprint size is 68µm×78µm.

If the same footprint need not be kept (*i.e.* source and drain electrodes need not be at opposite sides of the plate), further optimization can be performed. Figure 4.10 shows the 3rd generation 4T relay design. By keeping the source and drain regions to one side of the plate, A_{GS} and A_{GD} can be further minimized, and A_{GB} maximized. Note that the body (fixed W electrode) needs to be symmetric. Dimple and channel still need to be formed at the side of the plate opposite to the source/drain to act as a "stopper" to ensure parallel-plate actuation. (Without a "stopper", the plate could tilt or even collapse on one side.)

Figure 4.11 compares capacitances associated with the 1st, 2nd and 3rd generation relay designs, calculated based on the parallel-plate capacitor model:

$$C_{GB} = \frac{\varepsilon_0 A_{GB}}{\left(g_0 + \left(\frac{T_{ox}}{\kappa_{ox}}\right) - z\right)} \tag{4.3.2.1}$$

$$C_{GS} = \frac{\varepsilon_0 A_{GS}}{\left(g_0 + \left(\frac{T_{ox}}{\kappa_{ox}}\right) - z\right)} \tag{4.3.2.2}$$

$$C_{GD} = \frac{\varepsilon_0 A_{GD}}{\left(g_0 + \left(\frac{T_{ox}}{\kappa_{ox}}\right) - z\right)} \quad (4.3.2.3)$$

$$C_{GC} = \frac{\kappa_{ox}\varepsilon_0 A_{GC}}{T_{ox}} \quad (4.3.2.4)$$

$$C_{CB} = \frac{\varepsilon_0 A_{CB}}{(g_0 - z)} \quad (4.3.2.5)$$

where ε_0 is the permittivity of free space, κ_{ox} is the relative permittivity (dielectric) constant, g_0 is the as-fabricated actuation gap, and T_{ox} is gate oxide thickness. A_{GB}, A_{GS}, A_{GD}, A_{GC}, and A_{CB} denote the overlap areas between the two electrodes that forms the particular capacitance. The displacement $z = 0$ in the off-state and $z = g_d$ in the on-state, where g_d is the dimple gap.

In the 1st generation (original) design, the source and drain contribute 40% of the total capacitance associated with actuation. In the 2nd and 3rd generation designs, the source/drain contributions are reduced to 2.3% and 1.8% respectively. V_{GB} greatly dominates all parasitic components, resulting in dramatic improvement in gate control compared to the original design. The parasitic channel actuation effect is no longer observed in the 2nd generation design as C_{CB} is reduced to <3% of the original design, and completely eliminated in the 3rd generation design (no more channel-to-body overlap). The load capacitance of a relay circuit comprises the input capacitances of the relays in the next stage, which is dominated by C_{GB} in the off-state and C_{GC}+C_{GB} in the on-state. In the 3rd generation design, C_{GB} is nearly doubled while C_{GC} is reduced to less than a tenth of its value in the original design. Assuming complementary operation, where only half of the relays in the circuit are on at any one time, input capacitance is reduced to ~55% of that for the original design.

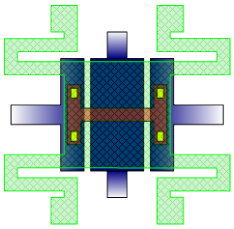
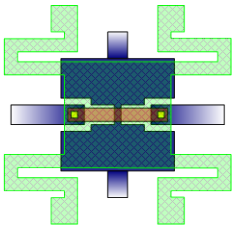
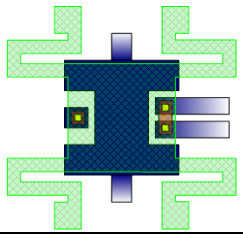
	1st Generation	2nd Generation	3rd Generation
Layout			
A_{SiGe}	810 μm^2	900 μm^2	900 μm^2
A_{dimple}	2 μm^2	1 μm^2	1 μm^2
$A_{dimpletotal}$	8 μm^2	2 μm^2	3 μm^2
g_0 (actuation gap)	130 nm	130 nm	130 nm
g_d (contact gap)	50 nm	50 nm	50 nm
T_{ox} (gate oxide thickness)	50 nm	50 nm	50 nm
A_{GB}	450 μm^2	754 μm^2	836 μm^2
C_{GB} (off-state)	29.39 fF	49.25 fF	54.60 fF
C_{GB} (on-state)	46.57 fF	78.03 fF	86.52 fF
A_{GD}	150 μm^2	9 μm^2	7.5 μm^2
C_{GD} (off-state)	9.80 fF	0.59 fF	0.49 fF
C_{GD} (on-state)	15.52 fF	0.93 fF	0.78 fF
A_{GS}	150 μm^2	9 μm^2	7.5 μm^2
C_{GS} (off-state)	9.80 fF	0.59 fF	0.49 fF
C_{GS} (on-state)	15.52 fF	0.93 fF	0.78 fF
$(C_{GS}+C_{GD})/C_{GB}$ (off-state)	0.667	0.024	0.018
$(C_{GS}+C_{GD})/C_{GB}$ (on-state)	0.667	0.024	0.018
C_{GB}/C_{TOTAL} (off-state)	0.600	0.977	0.982
C_{GB}/C_{TOTAL} (on-state)	0.600	0.977	0.982
$(C_{GS}+C_{GD})/C_{TOTAL}$ (off-state)	0.400	0.023	0.018
$(C_{GS}+C_{GD})/C_{TOTAL}$ (on-state)	0.400	0.023	0.018
A_{GC}	134 μm^2	58 μm^2	10 μm^2
C_{GC} (on-state)	213.56 fF	92.44 fF	15.94 fF
A_{CB}	76 μm^2	2 μm^2	0 μm^2
C_{CB} (on-state)	8.41 fF	0.22 fF	0 fF

Figure 4.11: Overlap areas and capacitances associated with the different 4T relay designs. 1st, 2nd, and 3rd generation designs are compared assuming the optimized technology parameters (Wafer #23 in Figure 4.8). In the 2nd and 3rd generation designs, V_{GB} greatly dominates all parasitic components, resulting in dramatic improvement in gate control compared to the 1st generation design.

4.4 Multi-Source/Drain (Output) Relay

Since a "stopper" structure is necessary at the opposite side of the source/drain to prevent tilting in the 3rd generation relay design, we can get extra functionality for free without any area or device behavior penalty by having an additional pair of source/drain (S/D) electrodes in place of the stopper. This gives rise to the concept of multi-source/drain relays. When placed in a circuit, this allows a single relay to have multiple outputs for different functions or voltage levels to reduce the device count in the circuit. Chapter 5 provides a more detailed discussion on the use of multi-source/drain relays in circuits.

4.4.1 Dual-Source/Drain (2-Output) Relay Design

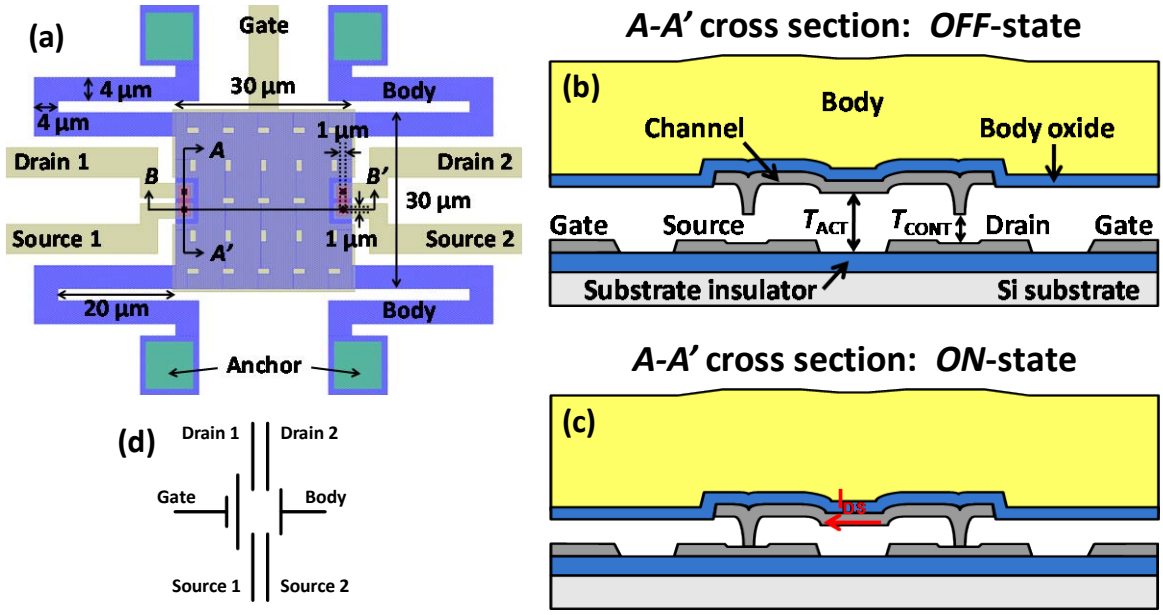


Figure 4.12: Illustrations of a dual-source/drain relay structure. **(a)** Layout view indicating the terminals and dimensions of relays fabricated in this work. **(b)** Cross sectional view along *A-A'* in the *OFF*-state and **(c)** in the *ON*-state. **(d)** Circuit symbol.

Figure 4.12 illustrates the structure and operation of a dual-source/drain (2-output) relay. In the off state, an air gap separates the channel and S/D electrodes so that no current flows. When the voltage difference between the gate and body electrodes is sufficiently large (*i.e.* when $V_{GB} \geq V_{PI}$, where V_{PI} is the "pull-in" voltage) the body electrode is actuated downward sufficiently to cause the channels – attached to the body electrode via an

intermediary oxide layer – to contact their respective S/D electrodes, so that current (I_D) can flow between the S/D electrodes. Figure 4.13 shows SEM images of a fabricated relay with a cross sectional view from focused ion beam (FIB) cut. Figure 4.14 shows measured I_D-V_G characteristics for standard devices (dimensions in Figure 4.12), demonstrating n-relay and p-relay operation by applying different body biasing. The curves from left side and right side source/drain pair overlap, indicating symmetric switching.

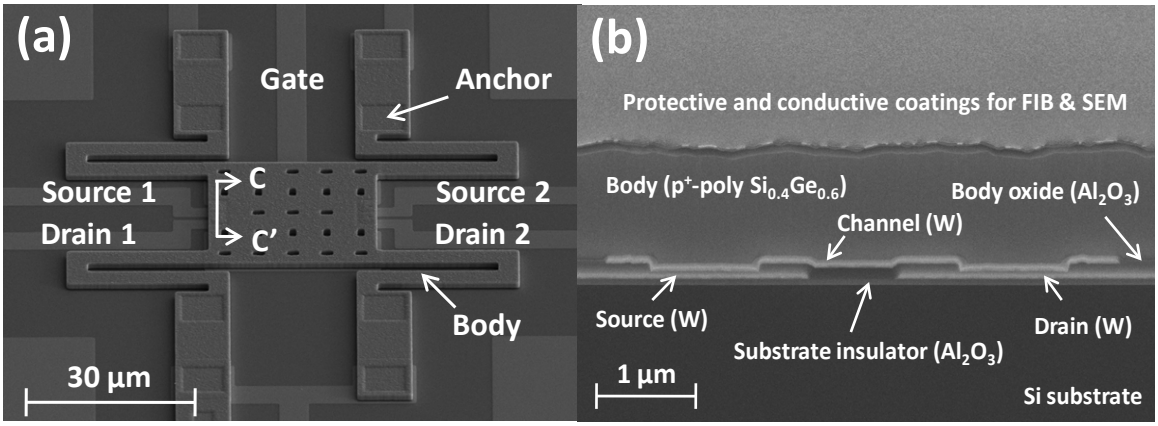


Figure 4.13: **(a)** Scanning electron micrograph (SEM) image of a dual-source/drain relay. **(b)** Cross-sectional SEM from focused ion beam (FIB) cut at C-C'.

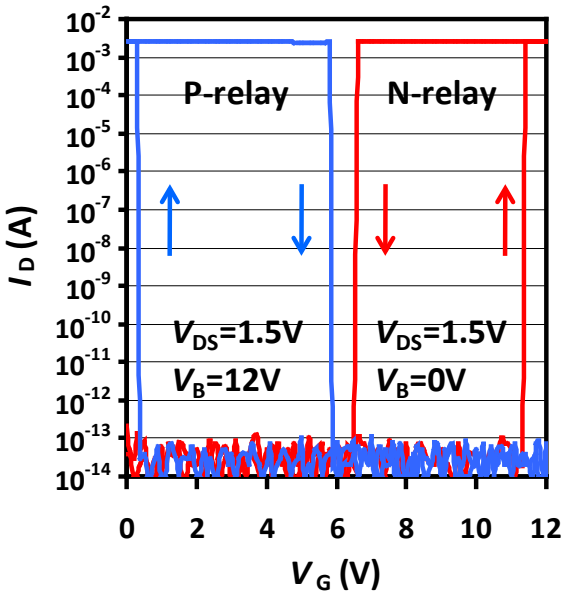


Figure 4.14: Measured I_{DS}-V_G characteristics of a dual-source/drain relay, with V_{DS} = 1.5 V. Operation mimicking that of either an n-channel or a p-channel MOSFET is seen by setting V_B at either 0 V or 12 V, respectively. Curves from both source/drain pairs sit on top of each other (same V_{PI} and V_{RL}), indicating symmetric switching.

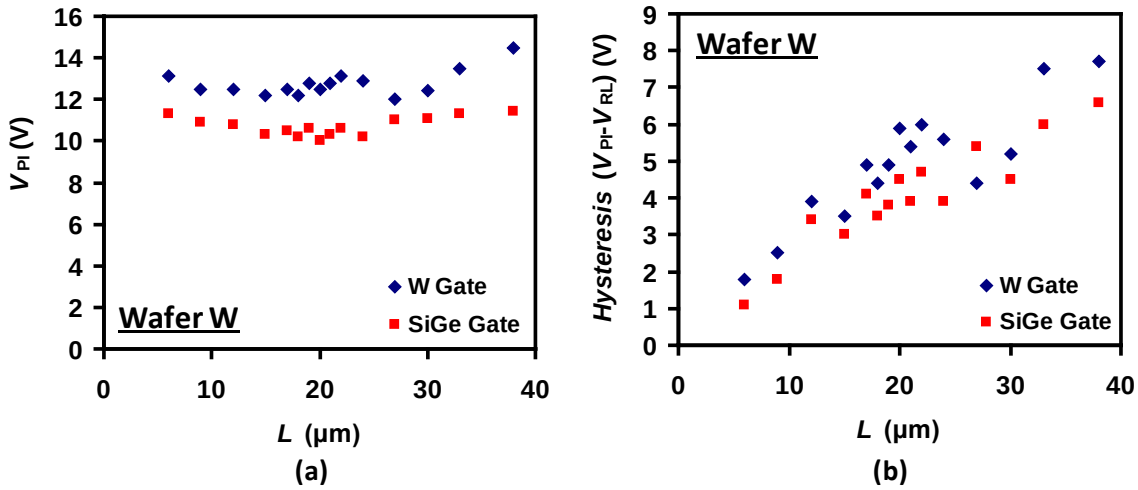


Figure 4.15: Comparison of W as gate *vs.* SiGe as gate for dual-source/drain relays of different flexure lengths (L), showing **(a)** V_{PI} and **(b)** hysteresis voltage. For the measurements, V_D = 1.5 V and V_S = V_B = 0 V. Biasing with SiGe as gate results in ~2 V lower V_{PI} and ~1 V lower hysteresis *vs.* W as gate.

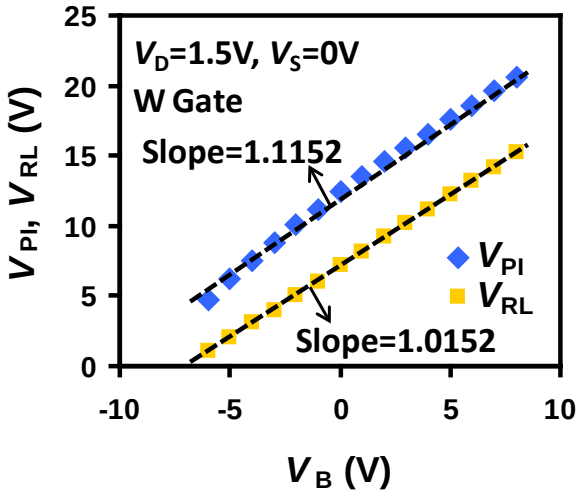


Figure 4.16: Dependence of the pull-in voltage (V_{PI}) and the release voltage (V_{RL}) on body bias (V_B) for a dual-source/drain relay biased using W as gate. V_D = 1.5 V, V_S = 0 V. For a given V_G, more negative V_B results in larger V_{GB} (larger electrostatic force) and so reduces V_{PI} and V_{RL}, and vice versa. Slopes are 1.1152 and 1.0152 for V_{PI} and V_{RL} respectively, which is close to the ideal case (slope = 1).

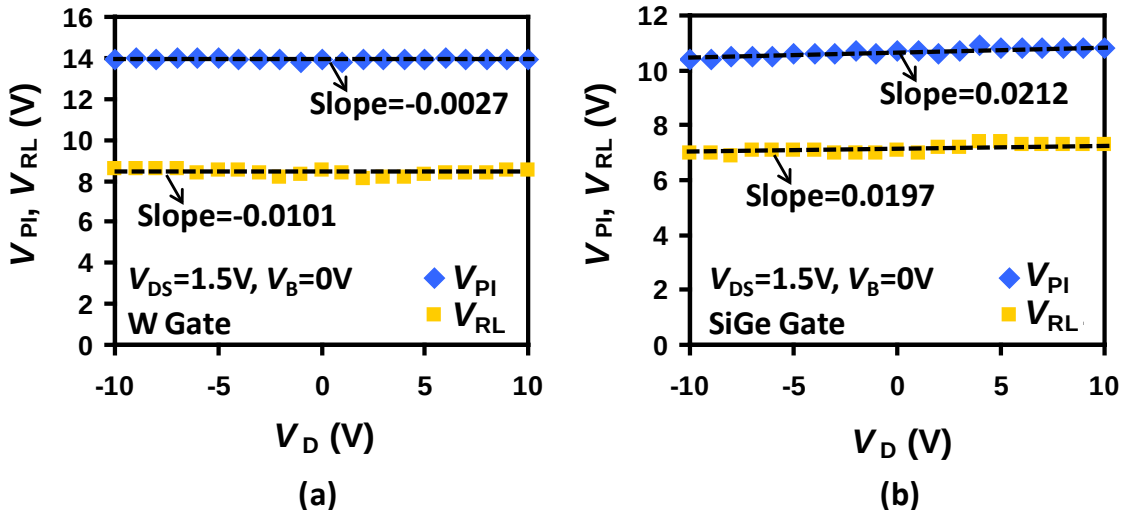


Figure 4.17: Dependence of the pull-in voltage (V_{PI}) and the release voltage (V_{RL}) on drain bias (V_D) for a dual-source/drain relay biased using **(a)** W as gate and **(b)** SiGe as gate. V_{DS} = 1.5 V and V_B = 0 V. Slopes are all close to ideal (slope = 0), indicating very minimal shift in V_{PI} and V_{RL} from parasitic electrostatic force between the gate and the source/drain.

Measurements in Figures 4.15-4.17 evaluate the effectiveness of gate control and parasitic actuation effects. When SiGe is biased as gate, V_{PI} is ~2 V lower and hysteresis ~1 V lower compared to using W as gate. Application of body bias results in corresponding reduction in switching voltages of equal magnitude. Actuation is independent of source/drain voltages. Therefore, this design is expected to be robust against parasitic electrostatic effects.

4.4.2 Quadruple-Source/Drain (4-Output) Relay Design

This concept of multiple-source/drain can be extended to relays with >2 source/drain pairs. As proof of concept, a quadruple-source/drain (4-output) relay is fabricated (Figure 4.18). Functionality is verified by measured I_D-V_G curves shown in Figure 4.19. Each drain is biased at different voltage levels to achieve different current levels. Actuation is found to be symmetric as switching voltages for all four source/drain pairs coincide. Note that symmetric switching is only achieved when the actuation plate is flat, stressing the increased importance of low strain gradient material for multi-source/drain designs.

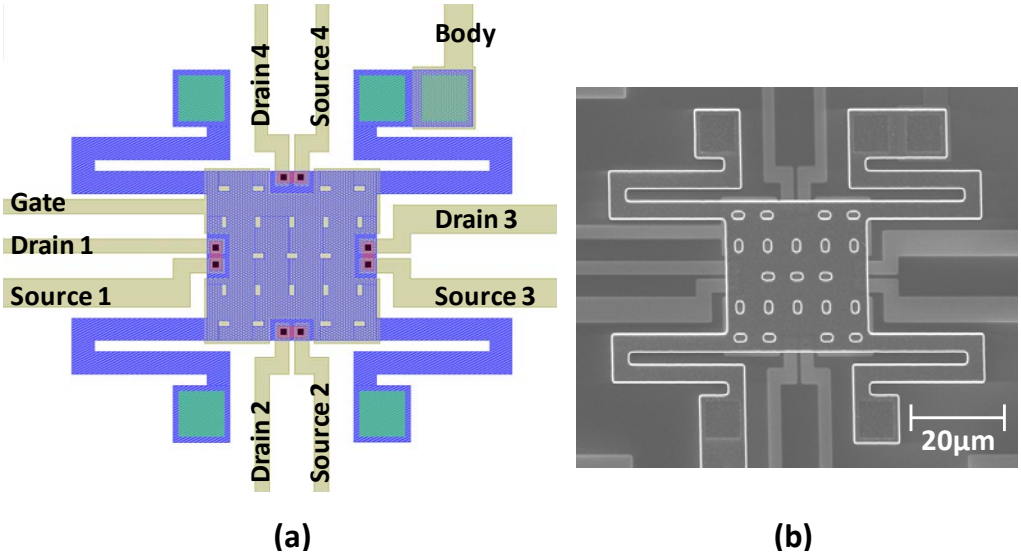


Figure 4.18: Illustrations of a quadruple-source/drain relay structure. **(a)** Layout view indicating the terminals. **(b)** Top-view scanning electron micrograph (SEM) image.

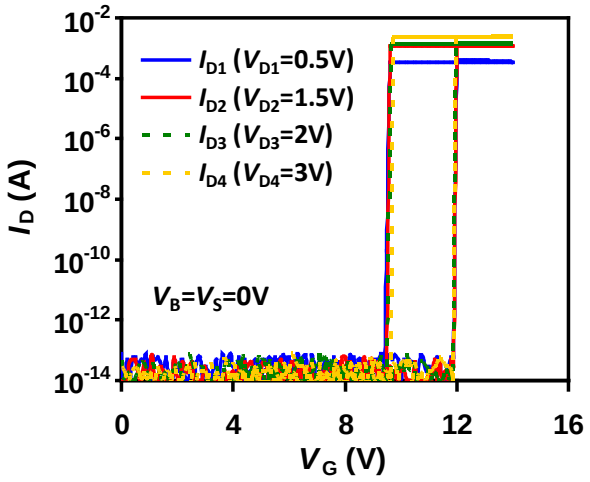


Figure 4.19: Measured I_{DS}-V_G characteristics of a quadruple-source/drain relay. Each drain are biased at different voltage levels (V_D = 0.5 V, 1.5 V, 2 V, 3 V). V_S = V_B = 0 V. Switching is simultaneous (same V_{PI}, V_{RL}) for all source/drain pairs, indicating that the four source/drain pairs are actuated symmetrically. Different current levels are seen, corresponding to different applied V_D's.

4.5 Multi-Gate (Input) Relay

4.5.1 Concept

Logic gate designs unique to relays can be achieved by taking advantage of the electrostatic force dependence on the actuation electrode area [6], [7]. For example, by carefully designing the beam dimensions, the number of driving input electrodes required to actuate a relay can be adjusted to implement two-input AND, OR [7], and NAND [6] gates. This design technique can be extended to other types of logic gates. This concept can be applied to the relay design in this work. The fixed electrode (used as gate) can be subdivided into multiple separate electrodes that can be independently biased. This way, the strength of electrostatic force (therefore actuation) can be controlled by the number of input electrodes (gate) that are driven. For a given process technology, there are several ways to design the relay to implement different logic functions:

(1) Subdivision of the fixed electrode.
Since electrostatic force depends on the total actuating electrode area, the fixed electrode can be replaced with smaller electrodes of equal area to accommodate multiple inputs - or of different areas to accommodate multiple inputs of varying weight (*i.e.* influence) - to implement different logic functions.

(2) Relay dimensions.
The flexure and plate dimensions determine the total electrostatic force needed to turn on the relay. For example, by adjusting flexure length (*i.e.* stiffness), a relay can be made to turn on with either one or two driven inputs.

(3) Body biasing.
Given a certain input voltage range, relays can be set to turn on when a specific numbers of input electrodes is driven by appropriately setting the level of body bias. In this manner, the same relay structure can be used to achieve different logic functions.

A possible drawback of the multi-input approach is scalability. When the fixed electrode is subdivided, there is a minimum separation distance between the electrodes. This distance is limited by surface leakage. In this work, the separation distance between electrodes is > 0.5 μm. As the size of the relay is scaled down, the amount of fixed electrode area lost due to electrode-to-electrode separation becomes a larger portion of the total area. As a result, actuation becomes less effective. Despite limited scalability, multi-electrode design potentially enables any logic function to be implemented with only two relays, which will be described later in Chapter 5. Therefore, overall area savings can be achieved from a system standpoint. In this work, dual-gate (2-input) relays are demonstrated. This concept can be extended to relays with >2 input electrodes [8].

4.5.2 Dual-Gate (2-Input) Relay Design

The fabricated dual-gate (2-input), dual-source/drain (2-output) relay structure is shown in Figure 4.20. Note that the gate electrodes have equal area and are inter-digitated to ensure that they have equal influence.

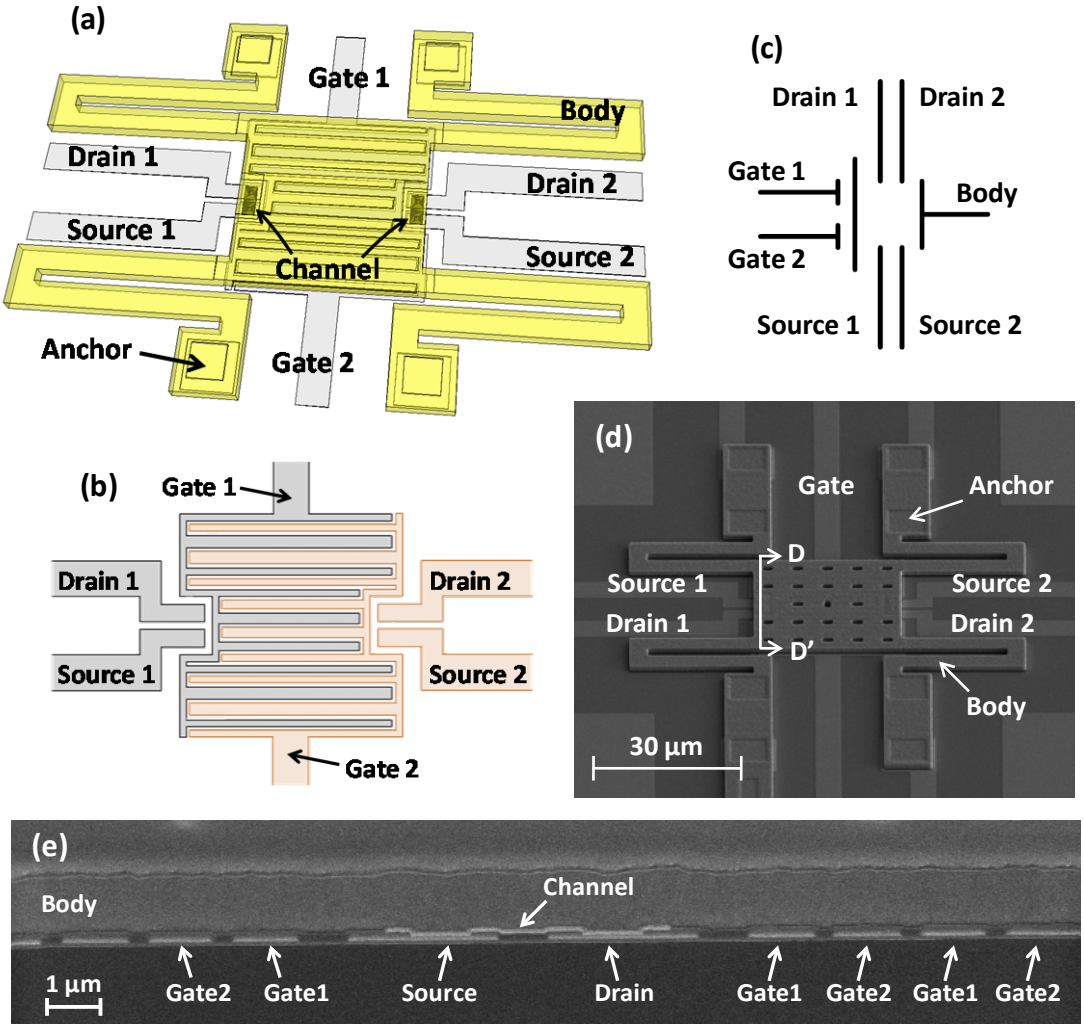


Figure 4.20: Schematic views of a dual-gate, dual-source/drain relay. **(a)** Isometric view. **(b)** Bottom electrode layout, showing inter-digitated gate electrodes to ensure that they have equal influence on the body. **(c)** Circuit symbol. **(d)** Scanning electron micrograph (SEM) image of a dual-gate, dual-source/drain relay. **(e)** Cross-sectional SEM from focused ion beam (FIB) cut at D-D'.

Figure 4.21 shows V_{PI} and V_{RL} of the dual-gate relay when $V_B = 0$ V. The first case, [0,1], is when only Gate 2 is driven, and the second case, [1,0], is when only Gate 1 is

driven. The non-driven gate is kept at 0 V. Switching voltages are equal for both cases, confirming Gate 1 and 2 have equal influence. The third case, [1,1], is when both gates are driven simultaneously to actuate the relay. As expected, the switching voltages are lower in this case due to stronger electrostatic force for the same voltage since the effective actuation area is doubled in this case. With this property, different levels of body bias can be applied to make the relay switch using either one gate or two gate electrodes for the same voltage. For example, an operating voltage of 12 V requires both gates to be driven to turn the relay on with V_B = 0 V (AND function). When V_B = -5 V is applied, the switching voltages are lowered, so that driving just one of the gates is sufficient to turn the relay on (OR function). In Figure 4.21, body biasing requirements to achieve AND and OR function are indicated for V_G = 8 V.

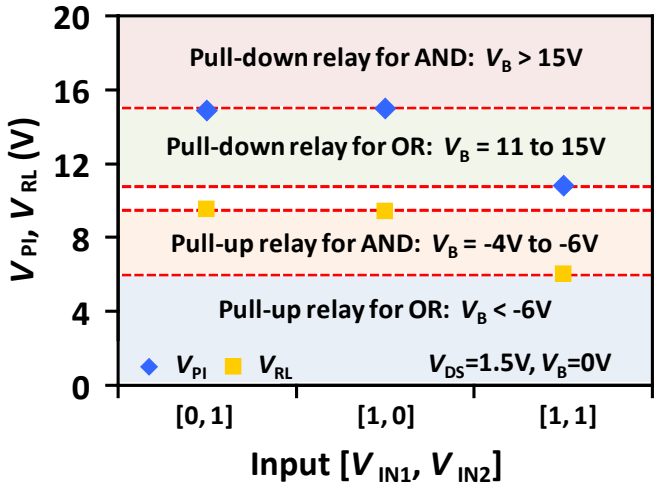


Figure 4.21: Measured pull-in (V_{PI}) and release (V_{RL}) voltages of a dual-gate relay, for various input combinations. "1" ≡ V_G. Body bias requirements to achieve AND and OR functions are indicated assuming V_G = 8 V.

4.6 Actuation Plate Designs

Without having to alter the established process, another knob we can play with is in the relay design itself, which is lithographically defined. While Equations 4.3.1.1-4.3.1.3 give insight on how to design for lower voltage, this is not straightforward since strain gradient effects tend to be dominant. The next few sections cover relay design optimization. Arrays of devices are fabricated to experimentally verify parameter-dependent trends.

Figures 4.22, 4.25, and 4.30 define the various parameters that are studied with respect to movable electrode, flexure, and the fixed electrode, respectively.

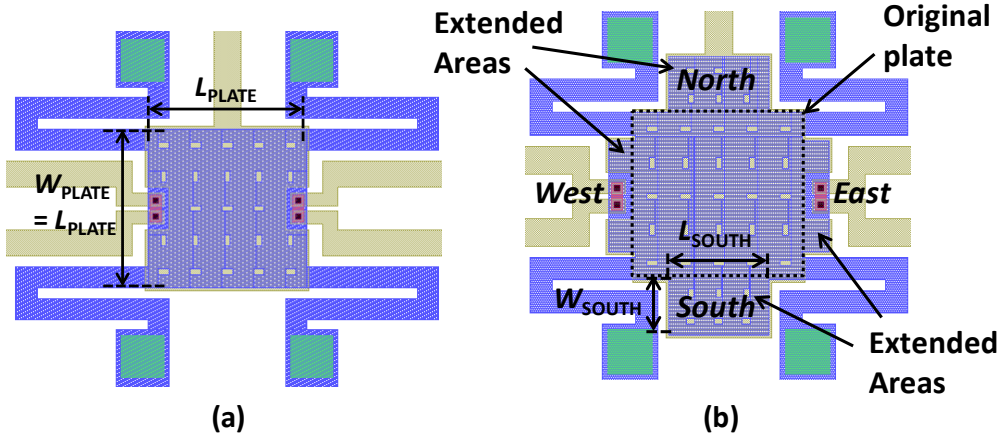
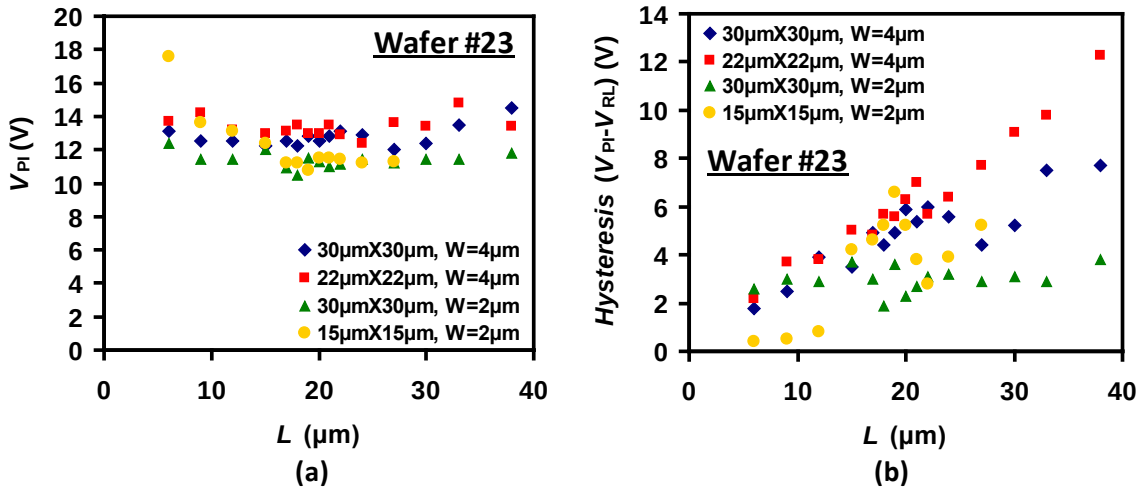


Figure 4.22: Plate design definitions. **(a)** Plate size for the original design is defined by plate width (W_{PLATE}) and length (L_{PLATE}). Plates are square in all designs presented here, *i.e.* $W_{PLATE} = L_{PLATE}$. **(b)** Extended actuation plate designs providesfor larger actuation area without any increase in the total device area, because they make use of empty spaces between anchors and flexures. The location of the extension is annotated by "North, South, East, West", or can be abbreviated by "NSEW". $L_{SOUTH} \times W_{SOUTH}$ denotes the area of the "South" extension. Similar convention applies for the other extensions. Each design is defined by the size of the original plate ($L_{PLATE} \times W_{PLATE}$) + total area of the extensions as a percentage of the original plate area.

4.6.1 Plate Sizes



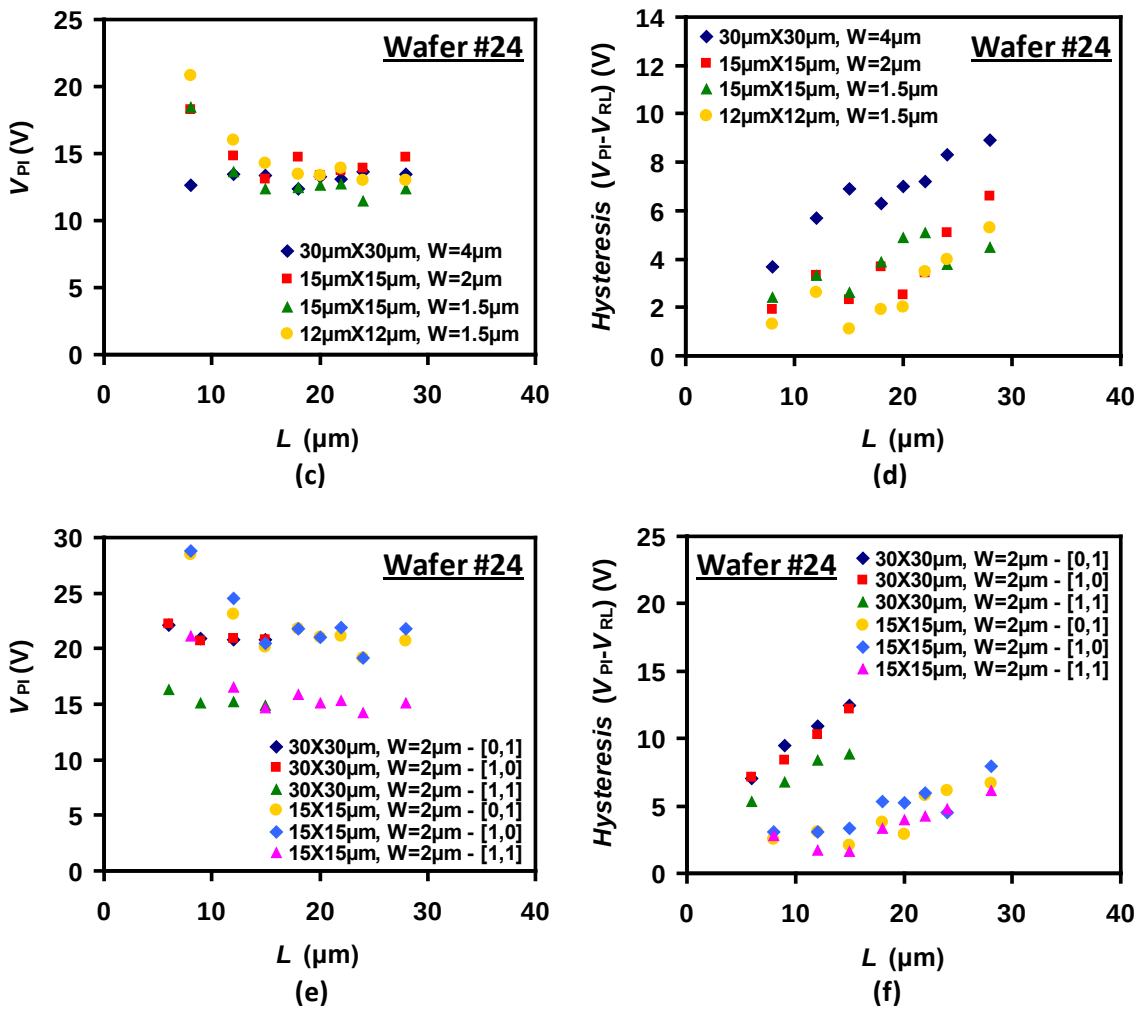


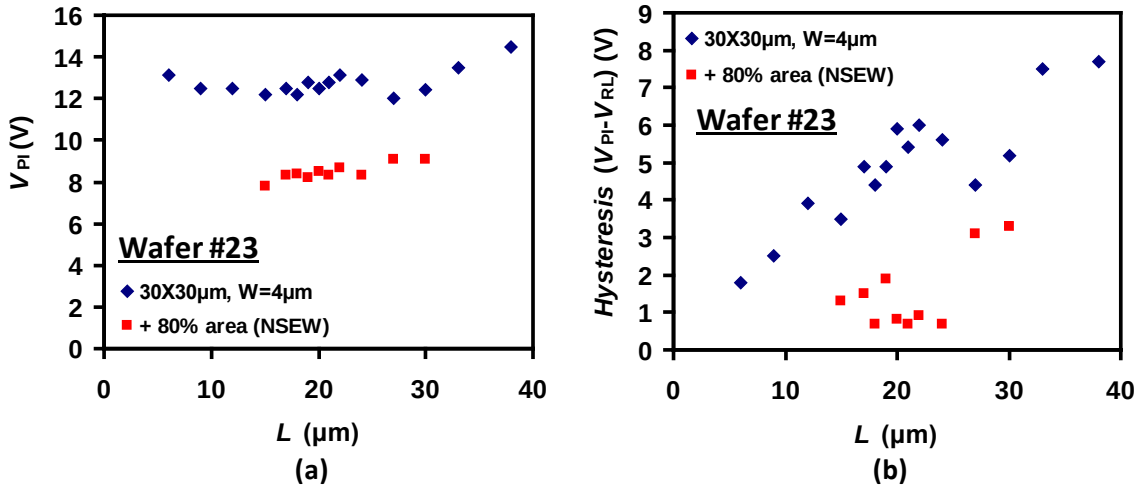
Figure 4.23: Effect of plate size on relay operating voltage. Relay V_{PI} and hysteresis voltage *vs.* flexure lengths (L) compared for relays with various plate sizes. **(a)**, **(b)** single-gate, dual-source/drain relays from Wafer #23. **(c)**, **(d)** single-gate, dual-source/drain relays from Wafer #24. **(e)**, **(f)** dual-gate, dual-source/drain relays from Wafer #24. For dual-gate relays, operation mode is denoted as [Gate1, Gate2] either at [0,1], [1,0], or [1,1]. "0"≡Gate is at 0 V. "1"≡Gate is swept from 0 V to V_{DD}. V_D = 1.5 V, V_S = V_B = 0 V. Tungsten fixed electrode is biased as gate. Process parameters of Wafers #23 and #24 are given in Figure 4.8.

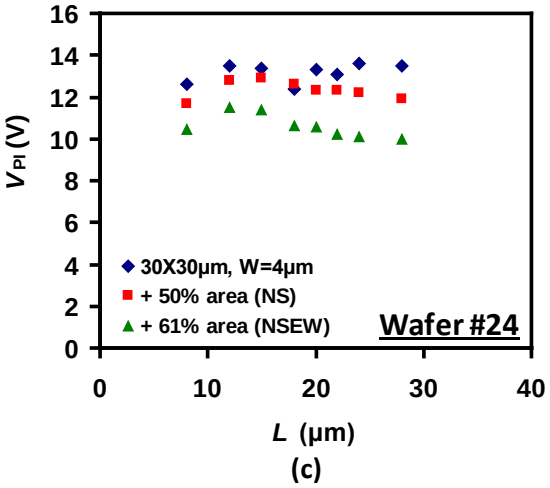
Figure 4.23 show the effect of plate size on relay switching voltages. Relays with plate sizes 30μm×30μm, 22μm×22μm, 15μm×15μm, and 12μm×12μm are fabricated. The scalability of the plate size depends on the smallest dimple size that can be reliably printed and etched. If the plate size is scaled continuously without scaling dimple size, the source/drain regions underneath the movable electrode will become a large percentage of the total actuation area, increasing parasitic electrostatic effects. Note that in some of these

cases, the flexure width (W) has been scaled together with the plate size. It is shown in the next section that flexure width has a significant influence. Thus, it is important to make comparisons fairly. For the same flexure width, scaling the plate size does not seem to impact V_{PI} too much, showing slight increase by ~0.5 V with smaller plate size. Despite reduced actuation area, smaller plate sizes help reduce warping due to strain gradient. Hysteresis voltages are comparable for the different plate sizes. In the dual-gate devices, 15μm×15μm devices show higher V_{PI} than 30μm×30μm devices only for short flexure lengths (L<12μm) and becomes comparable for longer flexure lengths. A significant reduction of hysteresis (by half) is observed in the 15μm×15μm devices. With reduced device area and potentially reduced hysteresis, smaller plate size is overall beneficial.

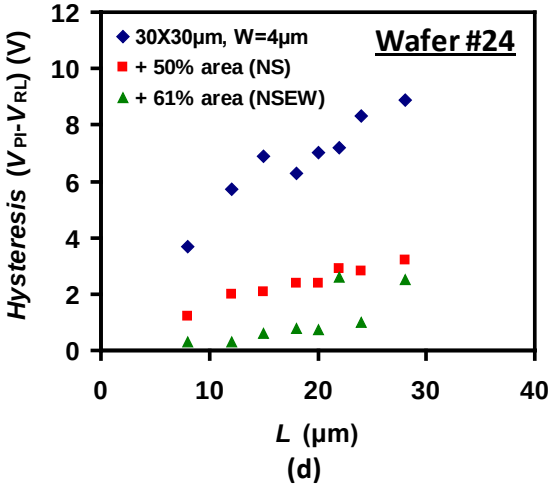
4.6.2 Extended Actuation Plate Area

Device footprint is set by the anchor placement (vertically) and the flexure lengths (horizontally). The standard design has plenty of empty space that can be filled with extensions of the movable plate. With extensions, reduction in V_{PI} is expected due to larger effective actuation area without increasing the total device footprint. A study of extended actuation plate effects is presented in Figure 4.24. In general, V_{PI} and hysteresis reduction is observed with larger extensions, but reliability degrades significantly when the extensions are too large. For example, 80% extra area to the 30μm×30μm devices results in ~4 V (30%) drop in V_{PI} and ~3-4 V drop (~70%) in hysteresis, but very poor yield and endurance, likely caused by strain gradient. Interferometer measurements confirm that negative strain gradient causes the extensions to bend downwards. With long extensions, this bending can be greater than the contact gap so that the channel touches the source/drain electrodes. Devices that are not stuck down are prone to stiction after a few cycles.

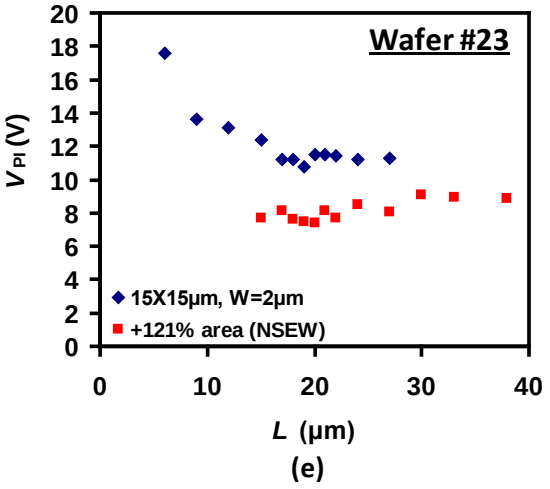



V_{PI} (V)
30X30μm, W=4μm
+ 50% area (NS)
+ 61% area (NSEW)
Wafer #24
L (μm)

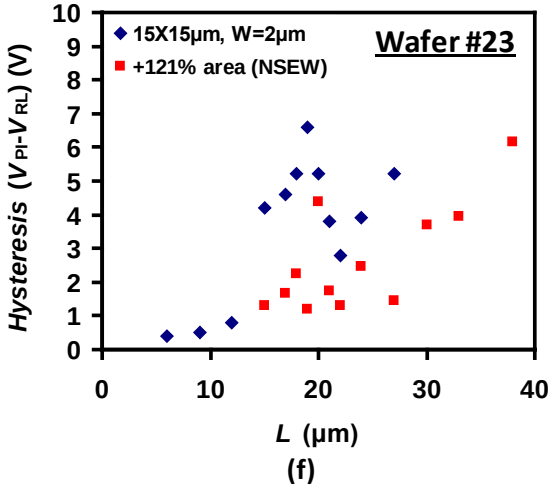
(c)


Hysteresis (V_{PI}-V_{RL}) (V)
30X30μm, W=4μm
+ 50% area (NS)
+ 61% area (NSEW)
Wafer #24
L (μm)

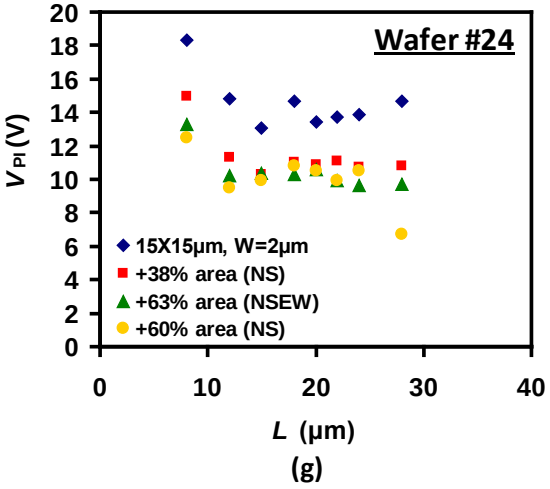
(d)


V_{PI} (V)
Wafer #23
15X15μm, W=2μm
+121% area (NSEW)
L (μm)

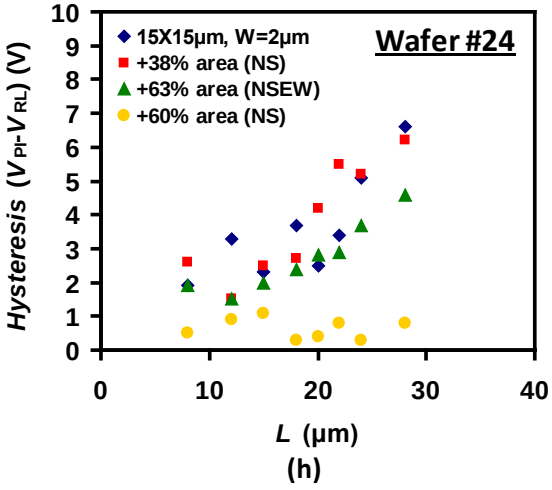
(e)


Hysteresis (V_{PI}-V_{RL}) (V)
15X15μm, W=2μm
+121% area (NSEW)
Wafer #23
L (μm)

(f)


V_{PI} (V)
Wafer #24
15X15μm, W=2μm
+38% area (NS)
+63% area (NSEW)
+60% area (NS)
L (μm)

(g)


Hysteresis (V_{PI}-V_{RL}) (V)
15X15μm, W=2μm
+38% area (NS)
+63% area (NSEW)
+60% area (NS)
Wafer #24
L (μm)

(h)

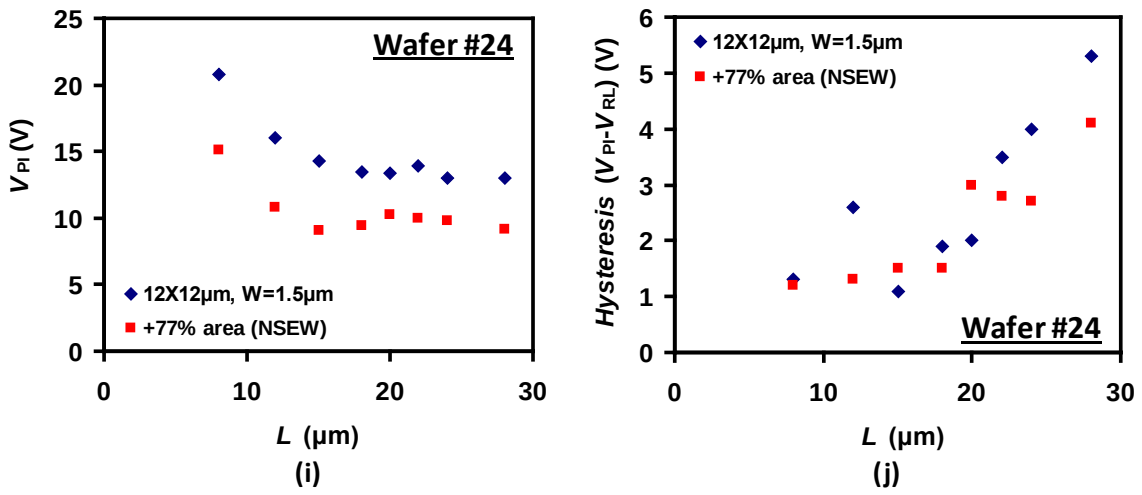


Figure 4.24: Effect of extended actuation plate area on relay operating voltage. Relay V_{PI} and hysteresis voltage *vs.* flexure lengths (L) compared for relays without and with extended actuation plate. **(a)**, **(b)** 30µm×30µm, W = 4 µm relays from Wafer #23. **(c)**, (**d)** 30µm×30µm, W = 4 µm relays from Wafer #24. **(e)**, **(f)** 15µm×15µm, W = 2 µm relays from Wafer #23. **(g)**, **(h)** 15µm×15µm, W = 2 µm relays from Wafer #24. **(i)**, **(j)** 12µm×12µm, W = 1.5 µm relays from Wafer #24. All relays are single-gate, dual-source/drain design. V_D = 1.5 V, V_S = V_B = 0 V. Tungsten fixed electrode is biased as gate. Process parameters of Wafers #23 and #24 are given in Figure 4.8.

Devices with ~60% extra plate area are found to be reliable, and appears to be the optimal point for maximum V_{PI} reduction *vs.* reliability tradeoff. 30µm×30µm and 15µm×15µm devices with 61% and 60% extra area respectively show >15% reduction in V_{PI} and hysteresis <1 V. For 12µm×12µm devices, 77% extra area still produce good yield and endurance. This shows that bending due to strain gradient is reduced with scaling.

4.7 Flexure Designs

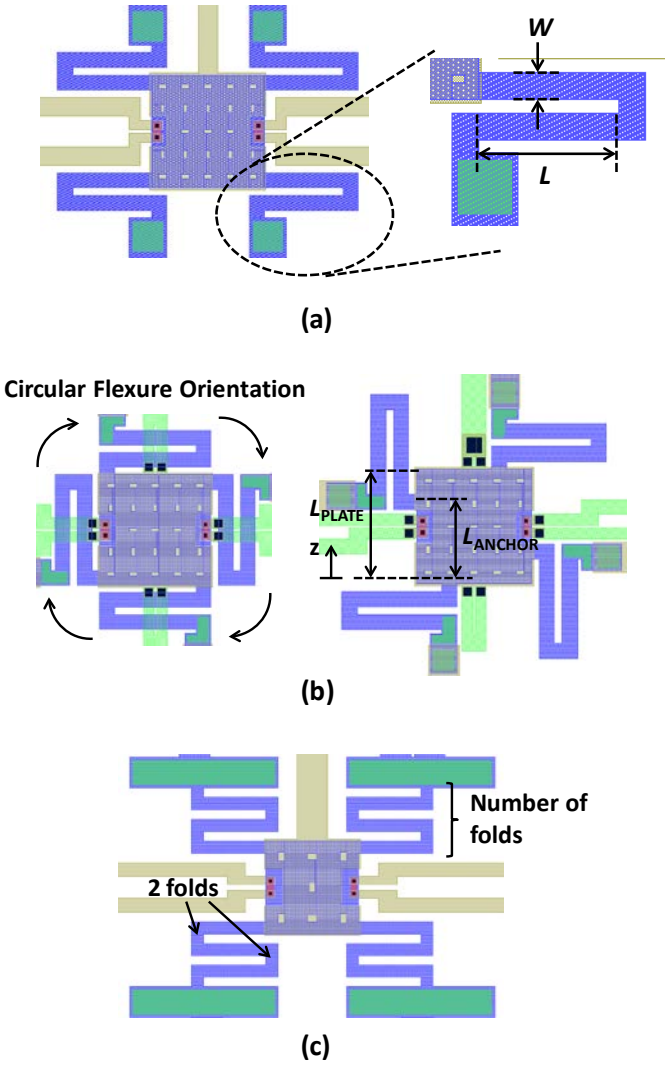


Figure 4.25: Flexure design parameter definition. (a) Definition of flexure width (W) and length (L). W is the width of the beam which is uniform throughout in all designs. L is defined for one beam section only, not the total beam length. (b) Circular flexure orientation in clockwise direction. In a clockwise orientation, distances (z) are measured from one edge of the plate also in a clockwise manner (annotated in the figure). Anchor position is defined by the ratio of the distance of the anchor from the edge of the plate (L_{ANCHOR}) to length of one side of the plate (L_{PLATE}), *i.e.* L_{ANCHOR} / L_{PLATE}. The plates for all designs are square, i.e. L_{PLATE} is equal for all 4 sides. (c) Definition of number of folds. A relay with 2-fold flexure is shown in (c) *vs.* standard 1-fold flexure in (a) and (b).

4.7.1 Flexure Length

Figure 4.26 shows the effect of flexure lengths (L) on the switching voltages. For the standard 30μm×30μm device, V_{PI} is found to be relatively insensitive to flexure length, while V_{RL} drops with increasing L, widening hysteresis. Since turn-on is brought about by electrostatic force while turn-off is brought about by spring restoring force, L has more significant impact on V_{RL}. V_{PI} starts to increase when flexure lengths are long (L>30μm) due to increased warping from strain gradient.

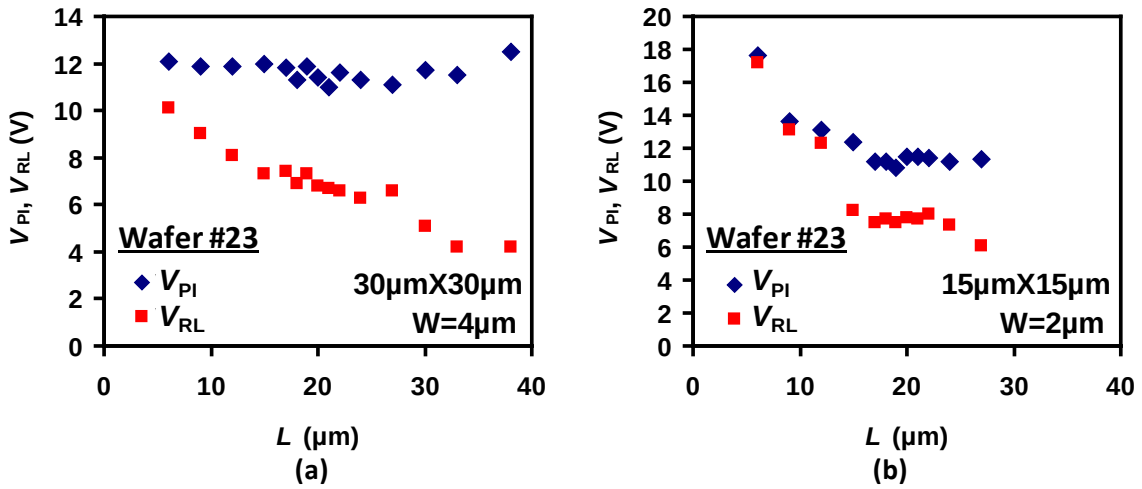


Figure 4.26: Effect of flexure length (L) on relay operating voltage. **(a)** V_{PI} and V_{RL} of 30μm×30μm, W = 4 μm relays. **(b)** V_{PI} and V_{RL} of 15μm×15μm, W = 2 μm relays. All relays are single-gate, dual-source/drain design from Wafer #23. V_D = 1.5 V, V_S = V_B = 0 V. Tungsten fixed electrode is biased as gate. Process parameters of Wafers #23 are given in Figure 4.8.

For the scaled devices (22μm×22μm and 15μm×15μm shown here), V_{PI} shows length dependence only for short flexure lengths (L<15μm). When the smaller relays turn on, spring restoring force is comparable to electrostatic force for the stiffer flexures to have an influence on V_{PI}, whereas electrostatic force is completely dominant for the large devices. For longer beams (L>15μm), V_{PI} becomes insensitive to L and hysteresis starts to widen as L gets longer, similarly as for the 30μm×30μm devices. L in the range from 15 μm to 20 μm appears to be optimal to keep V_{PI} and hysteresis low.

4.7.2 Flexure Width

The effect of flexure width (W) on switching voltages is studied in Figure 4.27. As W gets smaller, reductions in V_{PI} and hysteresis are observed. Take W = 5 μm as a

reference point. V_{PI} is reduced by ~1 V and ~2 V for W = 3 μm and W = 2 μm, respectively. Hysteresis is reduced by 1-2V and 2-3V, respectively. W = 2 μm is found to be the narrowest beam width that still allows for reliable operation. Despite large V_{PI} reduction by 3-4 V and very low hysteresis (<1 V), W = 1 μm devices show significantly poorer endurance because the flexures get too soft. Not only are they prone to stiction, the switching voltages drift considerably between cycles, indicating possible fatigue.

In the 15μm×15μm devices, W = 1.5 μm appears to be the minimum width limit at which devices still operate reliably. Similar to the 30μm×30μm devices, W = 1 μm devices have low V_{PI}, but show poor yield and endurance and are therefore not useful.

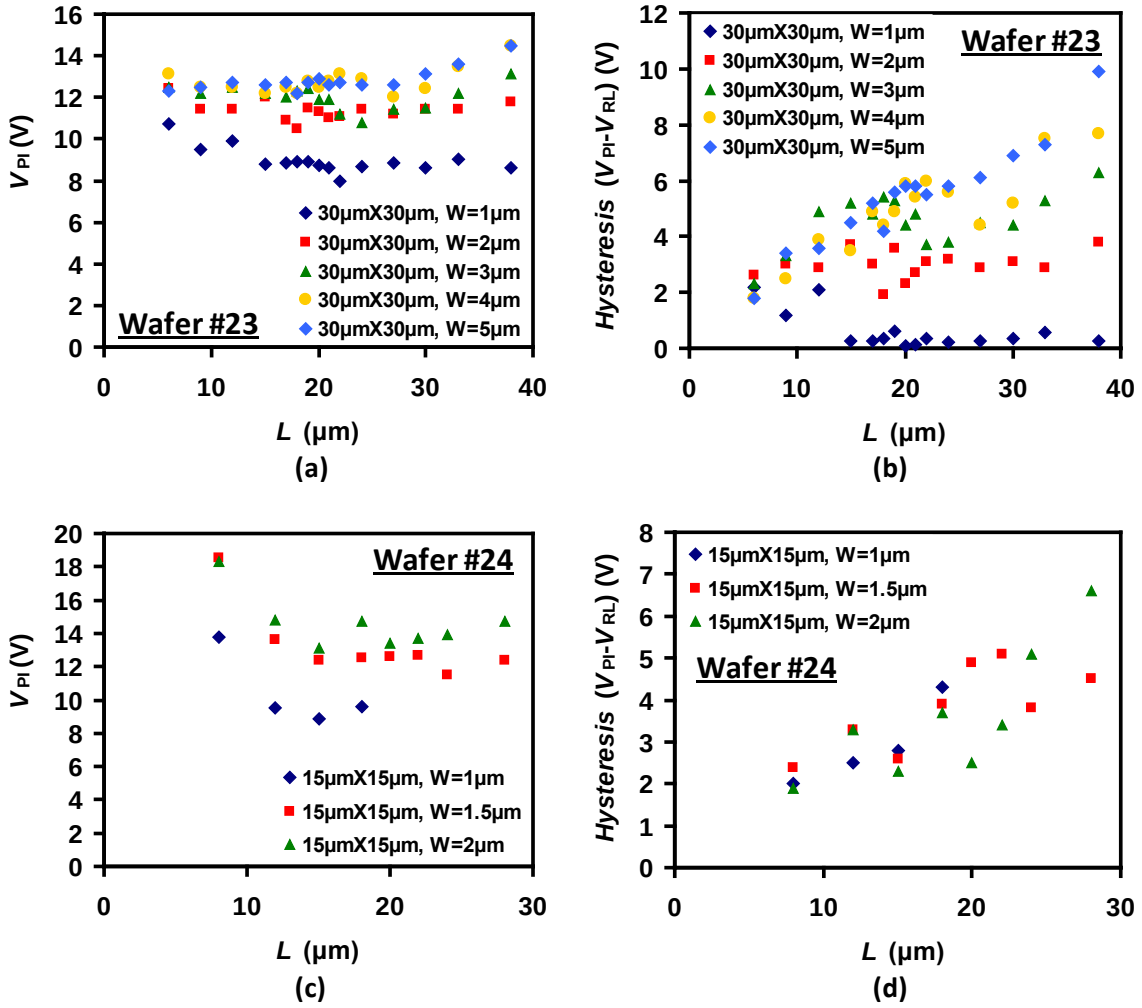


Figure 4.27: Effect of flexure width (W) on relay operating voltage. Relay V_{PI} and hysteresis voltage *vs.* flexure lengths (L) compared for various W. **(a)**, **(b)** 30μm×30μm relays from Wafer #23. **(c)**, **(d)** 15μm×15μm relays from Wafer #24. All relays are single-

gate, dual-source/drain design. V_D = 1.5 V, V_S = V_B = 0 V. Tungsten fixed electrode is biased as gate. Process parameters of Wafers #23 and #24 are given in Figure 4.8.

4.7.3 Flexure Orientation

In Figure 4.28, devices with flexures oriented circularly in a clockwise manner are studied. The anchor position is varied along the width of the plate, indicated by L_{ANCHOR}/L_{PLATE} ratio (described in Figure 4.25(b)). Significantly lower V_{PI} and hysteresis is found compared to the standard device. Devices anchored at L_{ANCHOR}/L_{PLATE} = 0.25 display lowest V_{PI}, but yield seems to be low. The highest yield is found for the devices anchored at the edge of the plate (L_{ANCHOR}/L_{PLATE}=0), with V_{PI} ~2 V lower than standard devices and hysteresis < 1V.

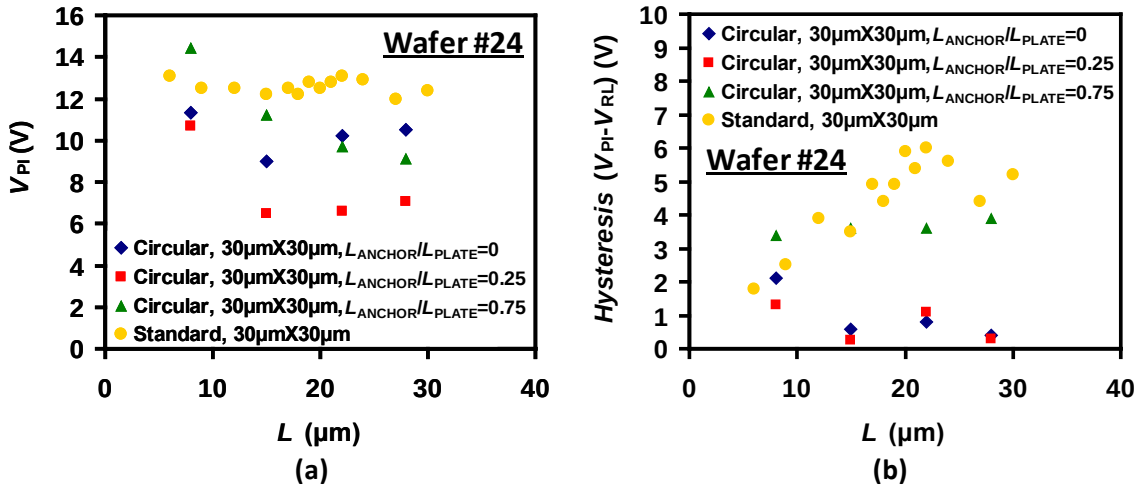


Figure 4.28: Effect of flexure orientation on relay operating voltage. Relay **(a)** V_{PI} and **(b)** hysteresis voltage *vs.* flexure lengths (L) compared for devices with standard and circular flexure orientations. In the circular case, flexures are oriented in a clockwise direction, with the anchor position denoted by L_{ANCHOR}/L_{PLATE} ratio (see Figure 4.25(b)). All relays are single-gate, dual-source/drain design with 30μm×30μm plate size from Wafer #24. V_D = 1.5 V, V_S = V_B = 0 V. Tungsten fixed electrode is biased as gate. Process parameters of Wafers #24 are given in Figure 4.8.

These designs however do not seem to be scalable. Small-sized devices (15μm×15μm) with small dimple size (0.5μm×0.5μm) do not switch (open circuit). It is likely that residual stress on the beam causes the plate to rotate slightly and the dimples are no longer aligned with the source/drain electrodes. Thus, no contact is made. The larger (30μm×30μm) devices switch successfully as they have larger dimple size (1μm×1μm) and

larger source/drain area, so that contact to source/drain can still be made even if the plate rotates slightly.

4.7.4 Number of Folds

By increasing the number of folds, the flexure length is effectively increased, so V_{PI} reduction is expected. Figure 4.29 shows the impact of adding another fold to the flexures. A 2-fold designs increases flexure length by 50% from the original 1-fold design. For the standard design, reduction in V_{PI} by >2V is observed, while hysteresis is reduced to close to 1V for the shorter beams (L<16μm). Significant improvement is not seen for circularly oriented flexure design.

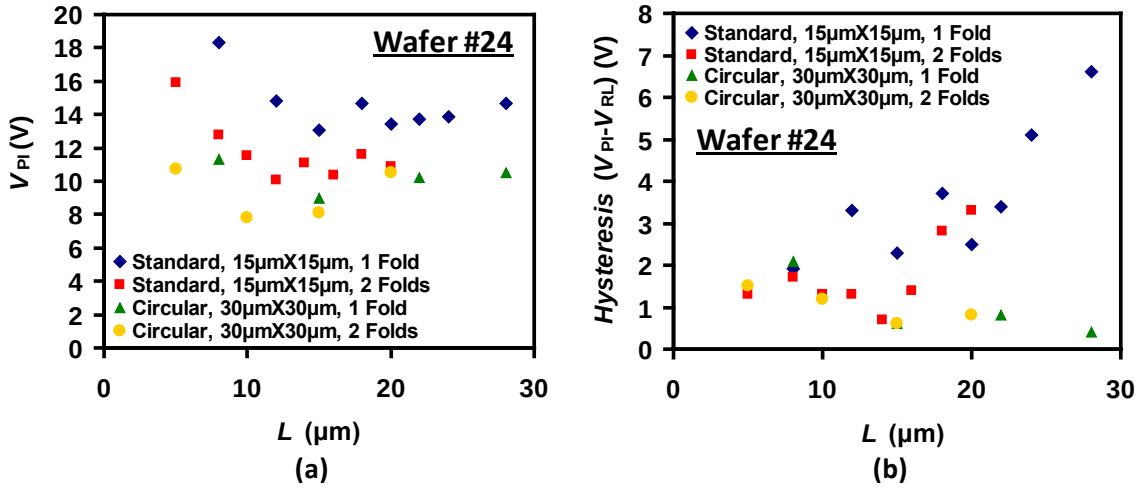


Figure 4.29: Effect of number of folds in the flexure on relay operating voltage. Relay **(a)** V_{PI} and **(b)** hysteresis voltage *vs.* flexure lengths (L) compared for devices with standard 1-fold *vs.* 2-fold flexures. Standard (15μm×15μm plate size) and circular (30μm×30μm plate size) flexure orientations cases are shown. In the circular case, flexures are oriented in a clockwise direction, with the anchor position at L_{ANCHOR}/L_{PLATE}=0 (see Figure 4.25(b)). All relays are single-gate, dual-source/drain design from Wafer #24. V_D = 1.5 V, V_S = V_B = 0 V. Tungsten fixed electrode is biased as gate. Process parameters of Wafers #24 are given in Figure 4.8.

4.8 Extended Fixed Electrode Area

Extending the fixed electrode area under the flexures effectively increases the actuation area (overlap between movable and fixed electrodes) for stronger electrostatic force. Figure 4.31 shows that a 3 V reduction in V_{PI} can be achieved when the whole length of the beam is covered on one side. Covering both sides of the flexures does not lower V_{PI} further. The side near the anchor is much stiffer, so that the additional electric force there does not have a significant impact on actuation.

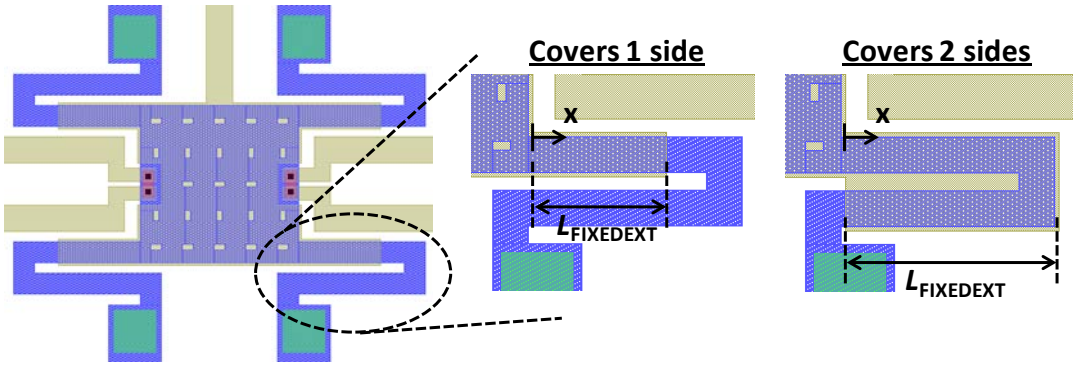


Figure 4.30: Designs with extended fixed (tungsten) electrode areas. The extensions are below the flexures, to provide for extra electrostatic force to actuate the relay. Extension length ($L_{FIXEDEXT}$) is measured from the original fixed electrode boundary. Designs that cover 1 side and 2 sides of the flexures are fabricated.

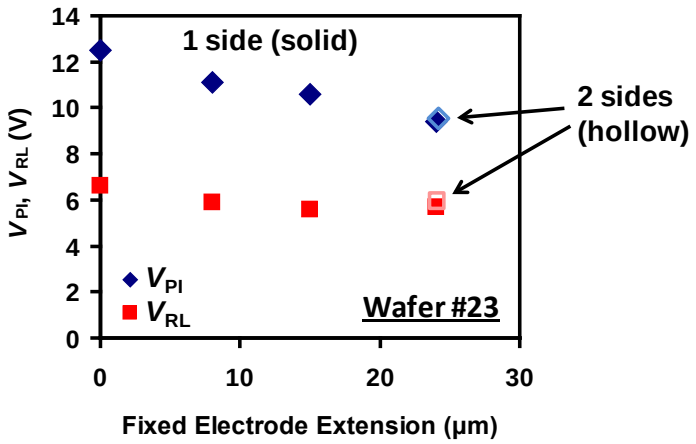


Figure 4.31: Effect of extended fixed (W) electrode on relay operating voltage. Relay V_{PI} and V_{RL} *vs.* fixed electrode extension length are shown. Solid data points show extension that only covers 1 side of the flexure. Hollow data points show extension that covers both sides of the flexure. All relays are single-gate, dual-source/drain design with 30μm×30μm plate size and L = 20 μm from Wafer #23. V_D = 1.5 V, V_S = V_B = 0 V. Tungsten fixed electrode is biased as gate. Process parameters of Wafers #23 are given in Figure 4.8.

4.9 References

[1] F. Chen, M. Spencer, R. Nathanael, C. Wang, H. Fariborzi, A. Gupta, H. Kam, V. Pott, J. Jeon, T.-J. K. Liu, D. Markovic, V. Stojanovic, and E. Alon, "Demonstration of integrated micro-electro-mechanical (MEM) switch circuits for VLSI applications," 2010 International Solid State Circuits Conference (San Francisco, California, USA), pp. 150-151, 2010.

[2] M. Spencer, F. Chen, C. Wang, R. Nathanael, H. Fariborzi, A. Gupta, H. Kam, V. Pott, J. Jeon, T.-J. K. Liu, D. Markovic, E. Alon, and V. Stojanovic, "Demonstration of integrated micro-electro-mechanical relay circuits for VLSI applications," IEEE Journal of Solid-State Circuits, Vol. 46, No. 1, pp. 308-320, 2011.

[3] H. Fariborzi, F. Chen, R. Nathanael, J. Jeon, T.-J. K. Liu, and V. Stojanovic, "Design and demonstration of micro-electro-mechanical relay multipliers," presented at the IEEE Asian Solid-State Circuits Conference (Jeju, Korea), November 2011.

[4] H. Fariborzi, M. Spencer, V. Karkare, J. Jeon, R. Nathanael, C. Wang, F. Chen, H. Kam, V. Pott, T.-J. K. Liu, E. Alon, V. Stojanovic, and D. Markovic, "Analysis and demonstration of MEM-relay power gating," presented at the 2010 Custom Integrated Circuits Conference (San Jose, California, USA), September 2010.

[5] H. Kam, “MOSFET replacement devices for energy-efficient digital integrated circuits,” Ph.D. Dissertation, University of California, Berkeley, 2009.

[6] K. Akarvardar, D. Elata, R. Parsa, G. C. Wan, K. Yoo, J. Provine, P. Peumans, R. T. Howe, and H.-S. P. Wong, “Design considerations for complementary nanoelectromechanical logic gates,” in Proc. International Electron Devices Meeting, pp. 299-302, 2007.

[7] A. Hirata, K. Machida, H. Kyuragi, and M. Maeda, “A electrostatic micromechanical switch for logic operation in multichip modules on Si,” Sensors and Actuators A, vol. 80, pp. 119-125, 2000.

[8] J. Jeon, L. Hutin, R. Jevtic, N. Liu, Y. Chen, R. Nathanael, W. Kwon, M. Spencer, E. Alon, B. Nikolic, and T.-J. K. Liu, "Multi-input relay design for more compact implementation of digital logic circuits," IEEE Electron Device Letters, Vol. 33, No. 2, pp. 281-283, 2012.

Chapter 5

Relay Based Combinational Logic Circuits

5.1 Introduction

Micro-relays recently have been investigated for digital integrated circuit (IC) applications because they are more robust against temperature variations and radiation than CMOS devices. Since relays have the ideal switching characteristics of zero off-state leakage and abrupt on/off switching behavior, which in principle allow for more aggressive voltage scaling, they are of interest for ultra-low-power IC applications as well [1], [2]. A highly reliable 4-Terminal relay nano-electro-mechanical (NEM) relay technology employing tungsten electrodes with endurance exceeding 10^9 on/off cycles was demonstrated in this dissertation, and subsequently was used to demonstrate digital logic, clocking, and memory circuits in [3]. A major advantage of the 4T relay design is that it allows for the pull-in and release voltages to be adjusted (post-process) via body biasing. This can be leveraged to achieve low-voltage operation, as well as to allow the same relay structure to be operated either as a pull-down device or as a pull-up device, mimicking n-channel MOSFET or p-channel MOSFET operation with the body biased at 0V or V_{DD} respectively. Thus, low-voltage complementary logic circuits can be implemented with 4T relays [4].

In this chapter, simple logic gates are demonstrated using fabricated relay devices described in previous chapters of this dissertation. The inverter is considered to be the basis for all logic circuits, as its operation principle is applicable to logic circuits in general. Thus, complementary relay inverter circuit characteristics are studied in depth. Body biasing schemes for a 4T relay inverter are compared with regard to the inverter voltage transfer characteristic, crowbar current, and noise margin. The optimal biasing scheme for achieving large noise margin and low-power operation is thereby identified.

Relays can be employed as direct replacements for transistors, as CMOS circuit design techniques are compatible with relays. Relay logic gates therefore are firstly demonstrated using CMOS-like design techniques. For the best performance, relay circuits need to be designed differently. Due to the large ratio between the mechanical (switching)

delay and the electrical (capacitive charging) delays of a relay, an optimized relay-based IC design should comprise a single-stage complex logic gate between latches, so that the time required to perform any digital logic operation is essentially one mechanical delay [2]. The properties of relays open up possibilities for unique relay-based circuit implementations. Unlike CMOS, relays are able to implement non-inverting logic since electrostatic force is ambipolar. Using multi-source/drain relays can increase functionality or significantly lower the number of devices required to implement a certain function. Finally, multi-input/multi-output designs provide the ability to implement complex logic circuits using only two devices, for ultimate compactness. These simple but significant circuit demonstrations pave the way for more complex circuit design using relays in the future.

5.2 Complementary Relay Inverter Circuit

5.2.1 Characteristics

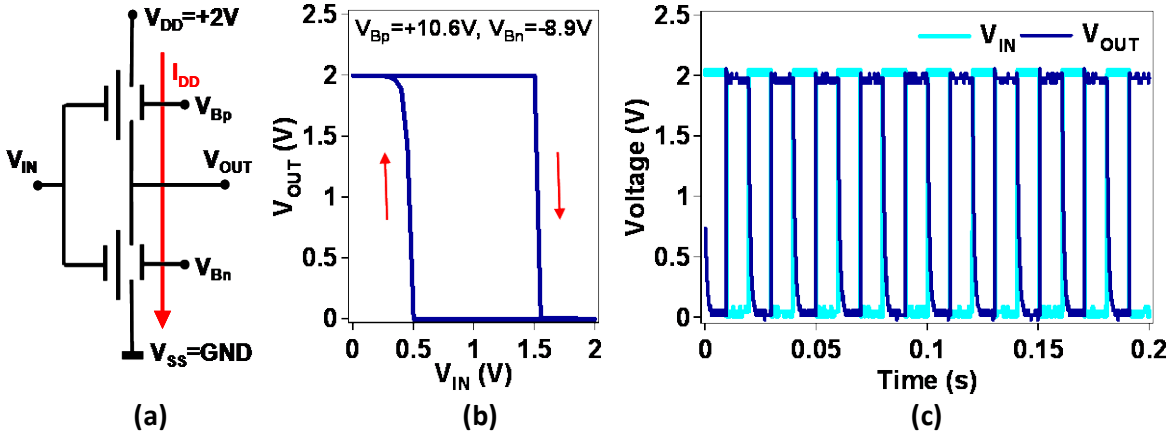


Figure 5.1: Demonstration of a complementary relay inverter. (a) Circuit schematic. Two relays are individually probed and connected externally. V_{DD} is set to 2 V. Body bias of the P-relay (V_{Bp}) and N-relay (V_{Bn}) is adjusted to achieve symmetric switching at low voltage. (b) Static inverter measurement with V_{Bp} = +10.6 V and V_{Bn} = -8.9 V. (c) Dynamic inverter measurement of the same relay circuit shown in (b), performed at a frequency of 50 Hz. V_{IN} is a square function, oscillating between 0 V and 2 V.

A relay can be biased to mimic either an n-channel MOSFET (turning on at sufficient positive V_{GS}) or a p-channel MOSFET (turning on at sufficient negative V_{GS}). Thus, an inverter circuit with zero static power dissipation can be formed by connecting two complementary 4T relays in series between the power supply and ground, similarly to the transistors in a CMOS inverter, as shown in Figure 5.1(a) to form an inverter circuit. The V_{PI} values for the two relays were carefully tuned via body biasing to achieve complementary switching characteristics for V_{DD} = 2 V, resulting in the nearly symmetric voltage transfer characteristics shown in Figure 5.1(b). Ideally, $V_{PIn} \geq V_{RLp}$ and $V_{PIp} \leq V_{RLn}$ to achieve abrupt high-to-low and low-to-high voltage transfer characteristics, respectively, and zero "crow-bar" current concomitantly. Dynamic inverter operation at 50 Hz is shown in Figure 5.1(c), for a square-wave input signal supplied by a waveform generator.

Note that in a CMOS inverter the transistors have gradual switching behavior (due to the fact that the sub-threshold swing is fundamentally limited to be no steeper than 60 mV/dec at room temperature) so that the CMOS inverter voltage transfer characteristic (VTC) shows a gradual transition between states, with non-zero "crow-bar current" (I_{DD}, flowing directly from the power supply to ground) in the transition region. In contrast, a relay inverter VTC can show abrupt transitions between states, with zero crow-bar current, if the relay switching voltages are tuned appropriately.

5.2.2 Body Biasing Schemes

Due to process-induced variations, it is difficult in practice to ensure that the N-relay V_{PI} and V_{RL} values exactly match the P-relay V_{RL} and V_{PI} values to achieve perfectly complementary switching, however. Figure 5.2 shows three possible situations: perfectly matched, $V_{PI,N} = V_{RL,P}$ and $V_{RL,N} = V_{PI,P}$ (Case A); lower N-relay switching voltages, $V_{PI,N} < V_{RL,P}$ and $V_{RL,N} < V_{PI,P}$ (Case B); and higher N-relay switching voltages, $V_{PI,N} > V_{RL,P}$ and $V_{RL,N} > V_{PI,P}$ (Case C). Along with the idealized relay I-V curves, the measured inverter voltage transfer characteristic (VTC) and measured crowbar current (I_{DD}) are shown for each case. (The N-relay and P-relay body biases were adjusted to achieve the particular switching-voltage alignment for each case). The measurements were made using an Agilent 4156C Semiconductor Parameter Analyzer. The voltage measurement unit in this instrument has a non-infinite input resistance, as evidenced by the higher current floor seen whenever the output node is connected to V_{DD} (*i.e.* whenever the P-relay is on).

In Case A, switching is perfectly complementary so that only one of the relays is in the on state for any value of V_{IN}. The VTC shows abrupt V_{OUT} transitions (high-to-low and low-to-high), and crowbar current is minimized.

In Case B, when V_{IN} is swept up from 0 V to V_{DD}, the N-relay switches on before the P-relay switches off. When V_{IN} is swept down from V_{DD} to 0 V, the P-relay switches

on before the N-relay switches off. Thus there are ranges of V_{IN} in which both relays are on. Within these ranges, V_{OUT} is at a value between 0V and V_{DD}, and crowbar current flows from V_{DD} to ground. (The actual value of V_{OUT} will depend on the relative strengths of the two relays).

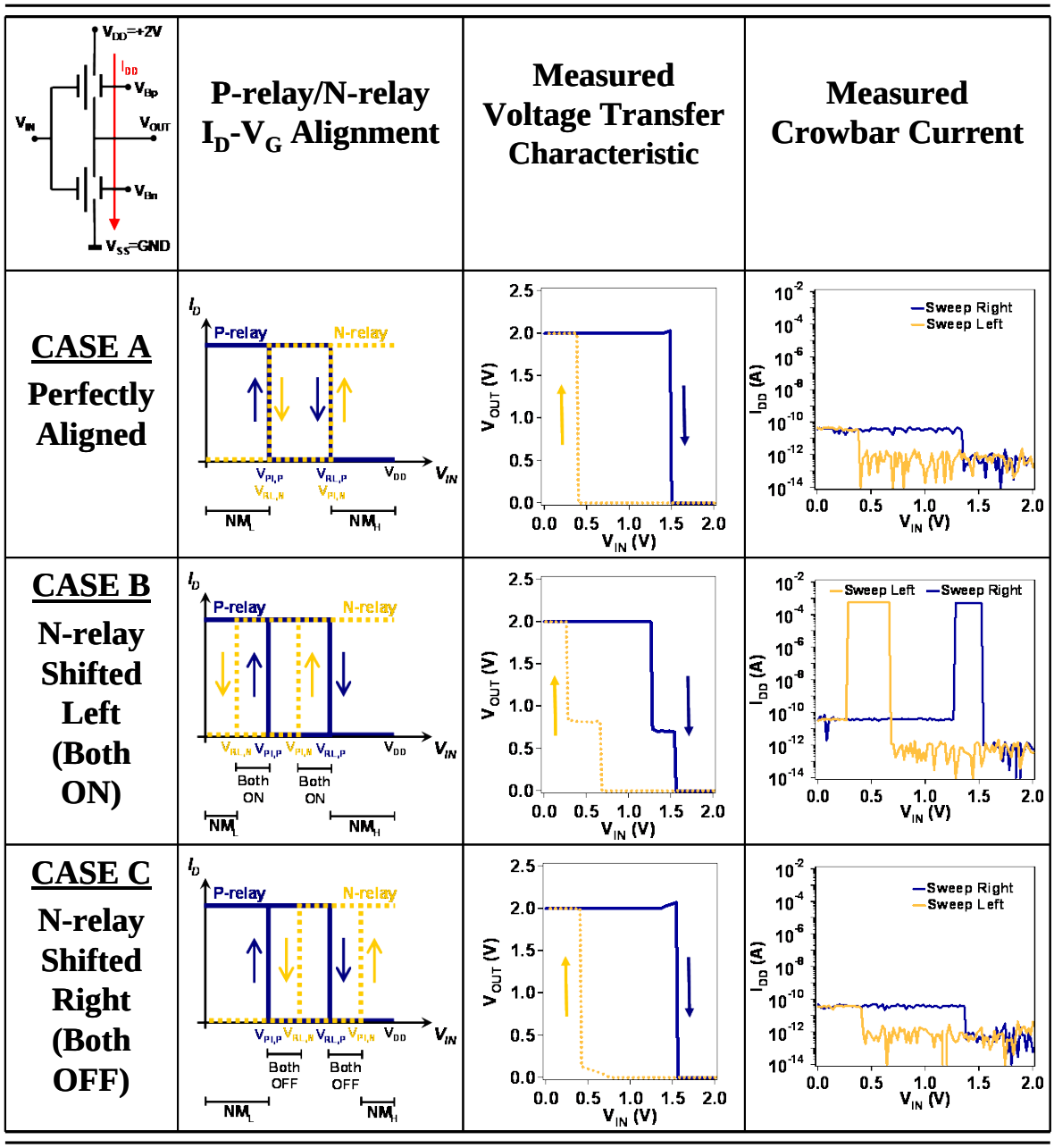


Figure 5.2: Relay inverter DC characteristics, for various body biasing schemes.

In actual circuit operation, the amount of time that both relays are on would be very short (less than the electrical charging/discharging delay time, t_{RC}), so that the extra power consumed during switching due the crowbar current should be negligible. Nevertheless, moderate relay on-state resistance (R_{ON}) is desirable in this case to avoid large current spikes. For digital logic applications, R_{ON} can be 10-100 kΩ because the throughput of an optimally designed relay-based circuit is limited by the mechanical pull-in time (10-100 ns) rather than by t_{RC} [2].

In Case C, when V_{IN} is swept up from 0V to V_{DD}, the P-relay switches off before the N-relay switches on. When V_{IN} is swept down from V_{DD} to 0V, the N-relay switches off before the P-relay switches on. Thus there are ranges of V_{IN} in which both relays are off. Within these ranges, fluctuations in V_{OUT} (due to residual leakage currents in the test setup) are seen in the VTC. In actual circuit operation, the amount of time that both relays are off would be very short, so that these voltage fluctuations should not adversely affect circuit operation. Since both relays are never on at the same time, crowbar current is minimized, similarly as for Case A.

Case A is ideal for minimizing the hysteresis in the VTC, but requires perfectly matched relays, which may be difficult to achieve and maintain in practice due to process-induced variations and device degradation under operation. To accommodate variations, it would seem best to bias the relays as in Case C (with the minimal misalignment needed to avoid the possibility of Case A occurring) to guarantee abrupt voltage transitions and zero crowbar current. In other words, for complementary logic and memory circuits, the relays should be biased such that $V_{PI,N} \geq V_{RL,P}$ and $V_{PI,P} \leq V_{RL,N}$.

5.2.3 Static Noise Margin

In a general digital integrated circuit, there is a need to both maintain the current logic state and to be able to flip the logic state. Noise margin is the smaller one of: (1) the maximum noise voltage that would not alter the logic state unintentionally and (2) the maximum noise voltage that would not prevent intentional flipping of the logic state. To illustrate this, consider the case where V_A = 0 V, V_B = 1 V. The logic state would flip if V_A reaches V_{PI} of the N-relay of Inverter 1 or V_B reaches the V_{PI} of the P-relay of Inverter 2, whichever comes first. For case (1), consider noise at V_A in the positive direction. If the magnitude of that noise is greater than the voltage required to flip the state, the logic will flip unintentionally. For case (2), consider noise at V_A in the negative direction. If the magnitude of that noise is greater than (V_{DD}-|Voltage required to flip the state|), we will not be able to flip the state even if we apply the maximum possible voltage (V_{DD}). The best possible noise margin occurs when all switching occurs at $V_{DD}/2$ (without hysteresis). When there is hysteresis, noise margin will be reduced and limited by case (2).

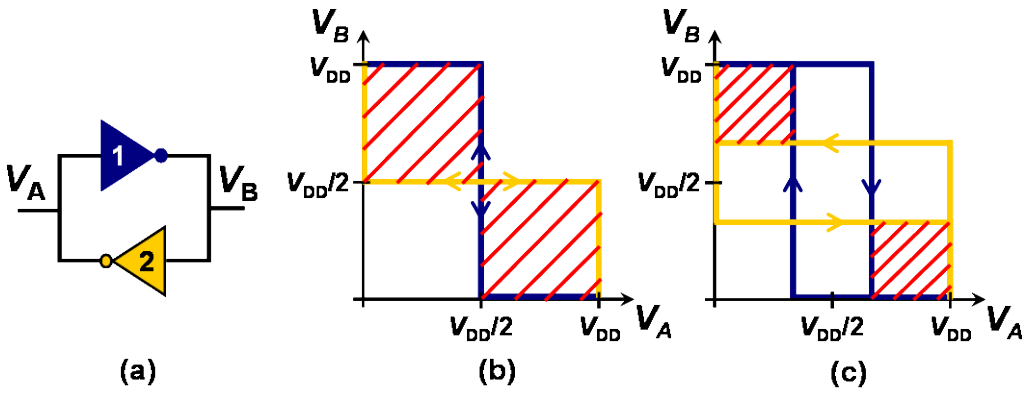


Figure 5.3: Static noise margin (SNM) of digital relay circuit. (a) Basic memory element (latch). (b) Maximum SNM possible is achieved when the voltage transfer characteristic (VTC) has zero hysteresis. (c) Noise margin is reduced with increasing VTC hysteresis.

For a CMOS inverter, the logic-low and logic-high noise margins (NML and NMH, respectively) are determined largely by the transistor threshold voltages: NML ~ $V_{TH,N}$; NMH ~ $|V_{TH,P}|$. For a relay inverter, NML is the smaller one of $V_{RL,N}$ and $V_{PI,P}$ and NMH is the smaller one of (V_{DD}-$V_{PI,N}$) and (V_{DD}-$V_{RL,P}$). Figure 5.3 illustrates how hysteresis in the VTC degrades the static noise margin (SNM). To achieve the maximum SNM, switching from high-to-low and low-to-high should be symmetric about $V_{DD}/2$, and hysteresis should be minimized. The results for the inverter circuit presented in this study are applicable to complementary logic circuits in general to fully utilize the benefits of the 4T relay technology.

5.3 CMOS-like Relay Circuits

5.3.1 Static Complementary Logic

A static complementary CMOS gate is essentially an extension of the complementary inverter circuit. It consists of a "pull-up network" and a "pull-down network" (Figure 5.4) [5]. When the output of the logic function is '1', the "pull-up" network provides a connection between the output and V_{DD}. Similarly, when the logic function outputs a '0', the "pull-down" network provides a connection between the output and GND. At steady state, only one of these networks will be on (conducting) at any given

time. Thus, the output will always have a connection to either V_{DD} or GND and static current flowing directly from V_{DD} to GND ideally should not exist.

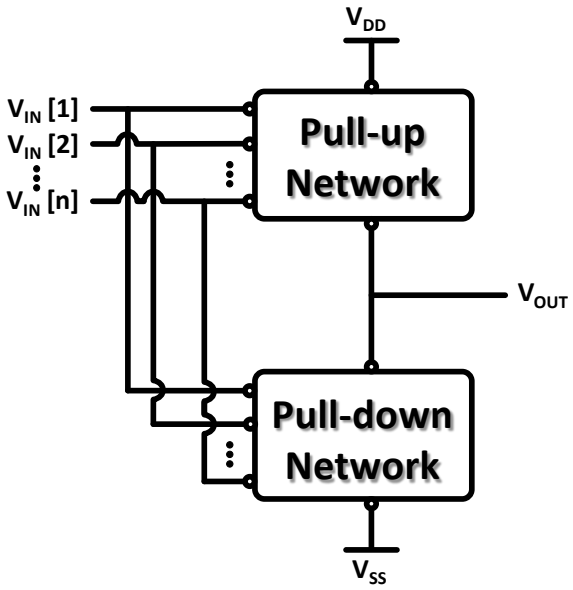


Figure 5.4: A generic static complementary CMOS logic circuit, consisting of "pull-up" network and a "pull-down" network. Complementary relay logic can be implemented in the same way.

Complementary relay logic gates can be implemented in a similar fashion with one major difference. In CMOS, the "pull-up" network is made up of PMOS transistors, while the "pull-down" network is made up of NMOS transistors. Consider Figure 5.5. An NMOS turns on when $V_{GS}=V_{TH,N}$, while a PMOS turns on when $|V_{GS}|=|V_{TH,P}|$. An NMOS is therefore able to discharge ("pull down") a node all the way down to 0 V, while it can only charge ("pull up") a node to V_{DD}-$V_{TH,N}$. A PMOS could charge a node all the way up to V_{DD}, but is only capable of discharging a node down to $|V_{TH,P}|$. As a result, complementary CMOS circuits are always inverting. Functions such as NAND, NOR and XNOR can be implemented in one stage, but functions like AND, OR, and XOR are formed by connecting an extra inverter to the inverting functions (thus incurring additional area and delay).

A relay, on the other hand, turns on when $|V_{GB}|=V_{PI}$, since electrostatic force is ambipolar. Hence, a relay can charge all the way to V_{DD} and discharge all the way to 0 V regardless of N-relay or P-relay operation mode. A complementary relay circuit can implement both inverting and non-inverting logic in a single stage. In fact, an optimal relay circuit should consist of a complex gate that performs all computation in a single stage (*ie.* only one mechanical delay).

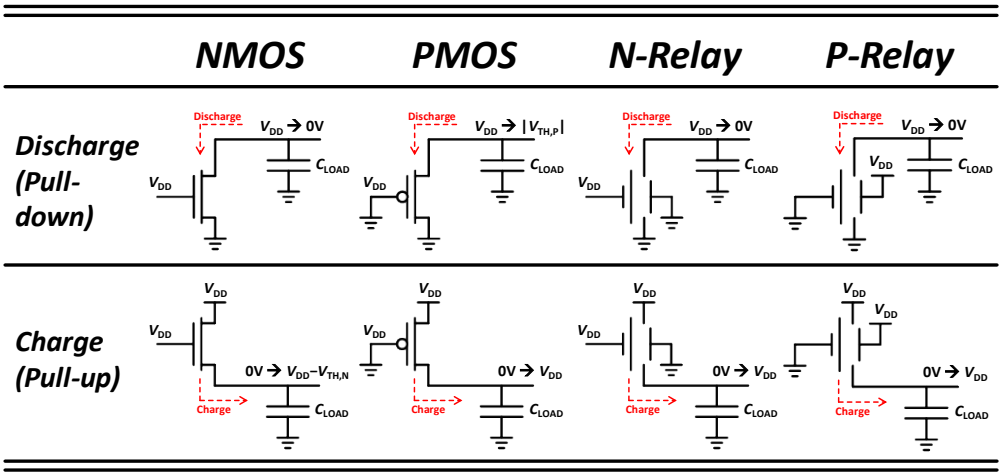
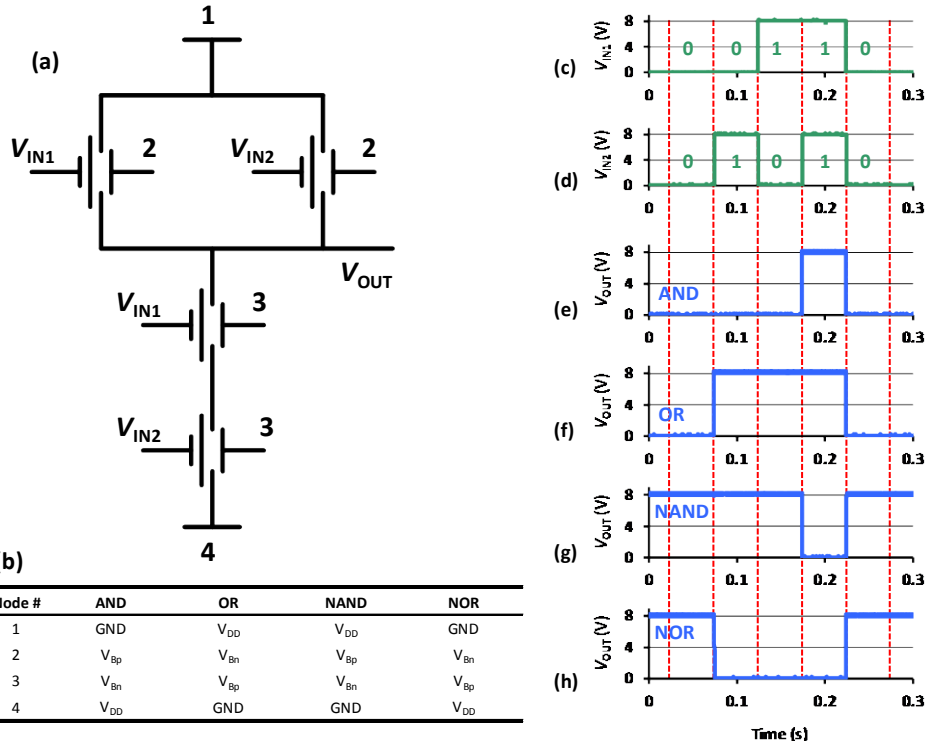


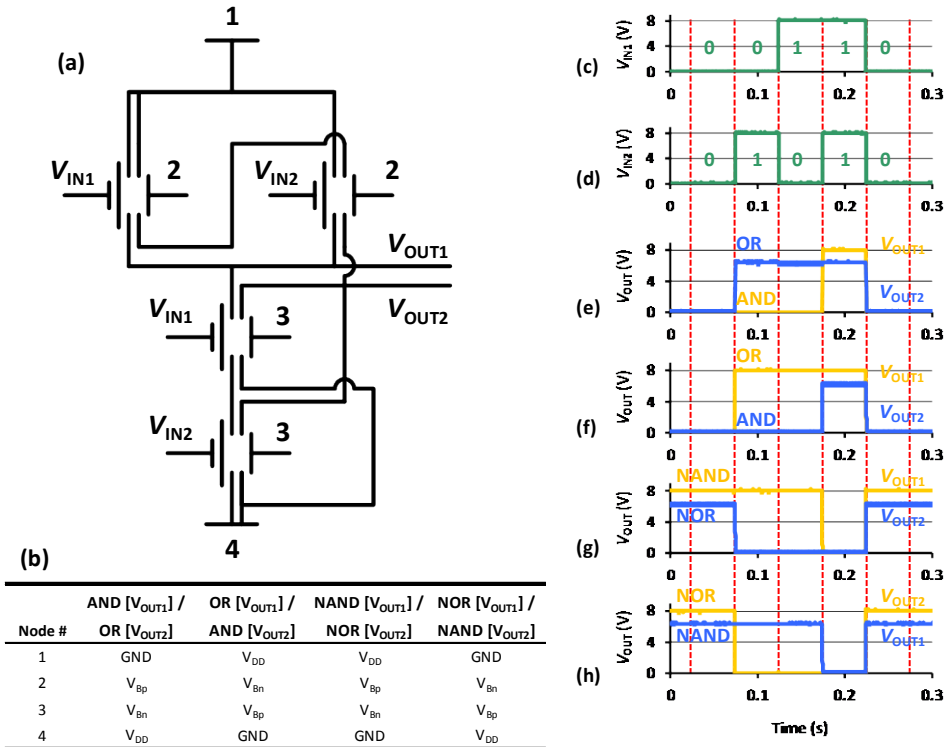
Figure 5.5: Comparison of NMOS, PMOS, N-relay, and P-relay when used as a "pull-down" and "pull-up" device.



Node #	AND	OR	NAND	NOR
1	GND	V_{DD}	V_{DD}	GND
2	V_{Bp}	V_{Bn}	V_{Bp}	V_{Bn}
3	V_{Bn}	V_{Bp}	V_{Bn}	V_{Bp}
4	V_{DD}	GND	GND	V_{DD}

Figure 5.6: A dynamically configurable complementary relay AND/OR/NAND/NOR gate implemented with 4T relays. **(a)** Circuit schematic. **(b)** Bias configurations. **(c)**, **(d)** Measured input waveforms. **(e)** Measured output waveforms for AND, **(f)** OR, **(g)** NAND, and **(h)** NOR. V_{DD} = 8 V, V_{Bp} = 13 V and V_{Bn} = -5V.

A dynamically configurable complementary relay logic gate using 4T relays is demonstrated in Figure 5.6. With 4 relays, AND/OR/NAND/NOR functions can be implemented by applying different biasing conditions. Thus, both inverting and non-inverting logic is achieved in a single stage. Dual-source/drain relay design can be utilized to add functionality to the logic gate by using the second pair of source/drain as a second output. The second output can implement either the complementary signal (to get differential output), the same signal with a different voltage level, or a different function altogether. As proof of concept, a relay logic gate that implements AND/OR, OR/AND, NAND/NOR, and NOR/NAND depending on bias configurations is demonstrated in Figure 5.7.



Node #	AND [V_{OUT1}] / OR [V_{OUT2}]	OR [V_{OUT1}] / AND [V_{OUT2}]	NAND [V_{OUT1}] / NOR [V_{OUT2}]	NOR [V_{OUT1}] / NAND [V_{OUT2}]
1	GND	V_{DD}	V_{DD}	GND
2	V_{Bp}	V_{Bn}	V_{Bp}	V_{Bn}
3	V_{Bn}	V_{Bp}	V_{Bn}	V_{Bp}
4	V_{DD}	GND	GND	V_{DD}

Figure 5.7: A dynamically configurable complementary relay AND/OR/NAND/NOR gate implemented with dual-source/drain relays to achieve multiple functionalities. **(a)** Circuit schematic. **(b)** Bias configurations. **(c)**, **(d)** Measured input waveforms. **(e)** Measured output waveforms for AND/OR, **(f)** OR/AND, **(g)** NAND/NOR, and **(h)** NOR/NAND. V_{DD} = 8 V, V_{Bp} = 13 V and V_{Bn} = -5V.

A second way to use the dual-source/drain relay design is to reduce the number of devices necessary to implement a certain function. For example, take a CMOS-like implementation of an XOR/XNOR gate using 4T relays (Figure 5.8). Although functional, an 8 relay implementation consumes large chip area. By employing dual-source/drain relays, the same circuit can be implemented with 4 relays (Figure 5.9). In this case, two 4T

relays with the same gate and body biases (colored red, green, blue and orange) are replaced with a single dual-source/drain relay, to reduce the device count by half.

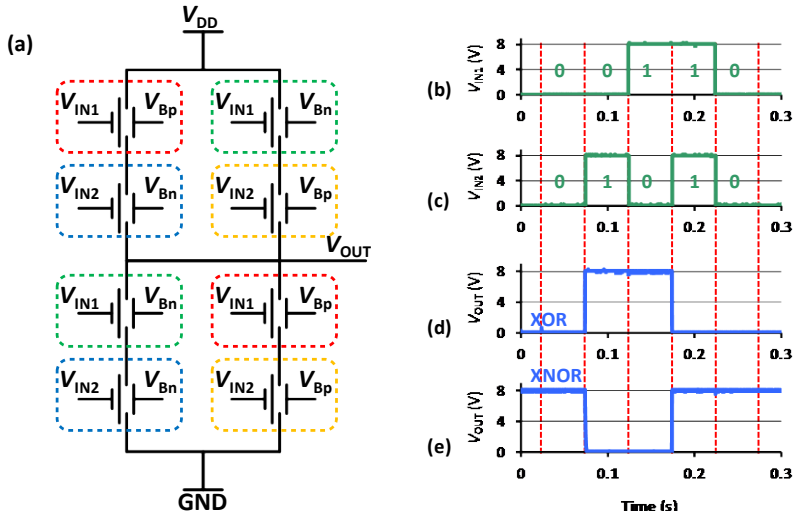


Figure 5.8: A complementary relay XOR/XNOR gate implemented with 4T relays. **(a)** Circuit schematic for XOR gate. XNOR function can be achieved by simply switching V_{DD} and GND. **(b)**, **(c)** Measured input waveforms. **(d)**, **(e)** Measured output waveforms. V_{DD} = 8 V, V_{Bp} = 13 V and V_{Bn} = -5 V. Annotated in different colors are relays that have identical gate and body biases.

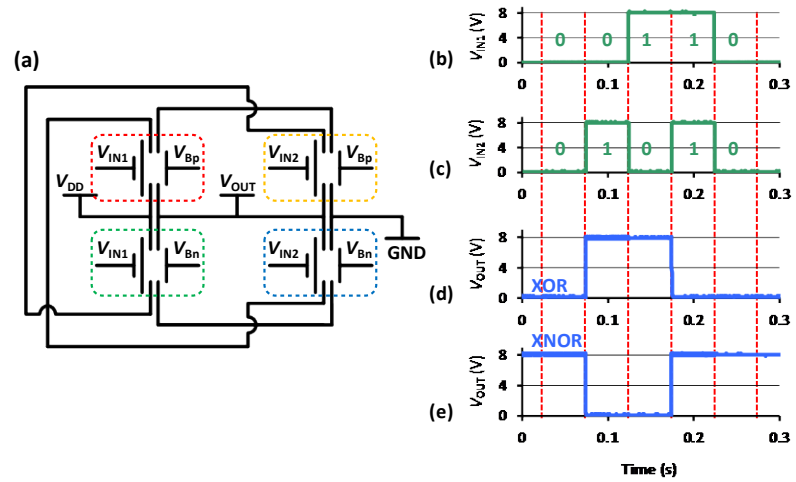
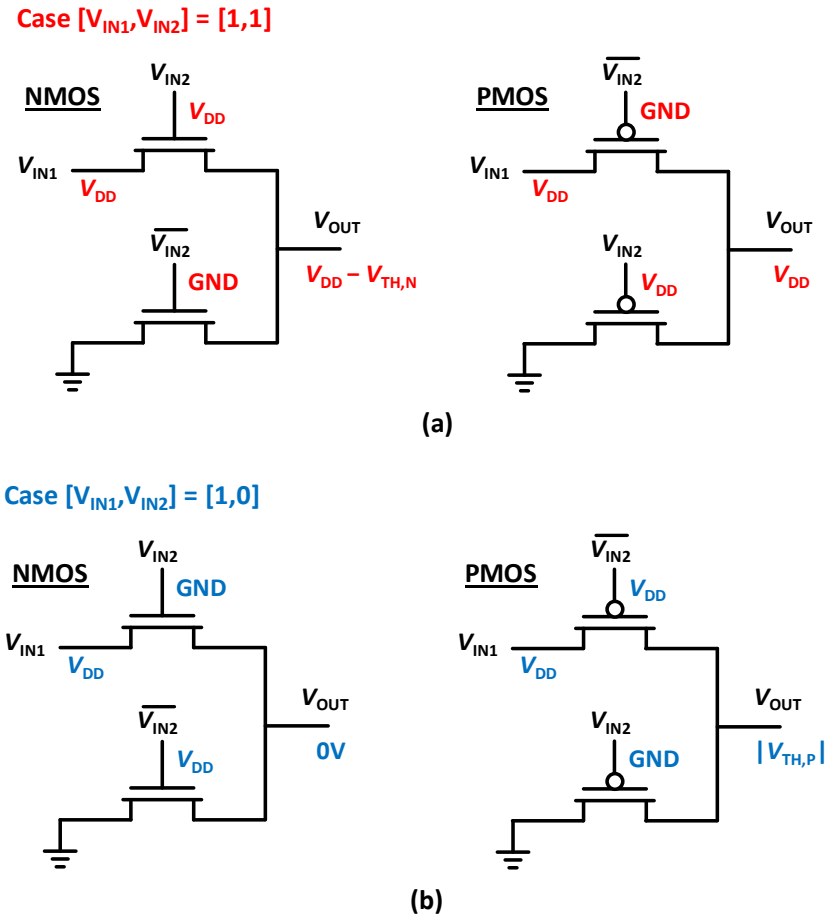


Figure 5.9: A complementary relay XOR/XNOR gate implemented with dual-source/drain relays to reduce device count for more compact logic. **(a)** Circuit schematic for XOR gate. XNOR function can be achieved by simply switching V_{DD} and GND. **(b)**, **(c)** Measured input waveforms. **(d)**, **(e)** Measured output waveforms. V_{DD} = 8 V, V_{Bp} = 13 V and V_{Bn} = -5 V. The two relays that have identical gate and body biases in Figure 5.7(a) are replaced with one dual-source/drain relay (annotated in the colors), reducing the device count by half.

5.3.2 Pass-Gate Logic

The pass-gate logic design in CMOS is an attractive alternative to complementary logic design to reduce device count. For example, in Figures 5.10(a),(b) an AND gate could be implemented with 2 transistors. (Note there is a requirement to invert the signal to get its complement that incurs another 2 transistors, for a total of 4). However, the drawback of an NMOS only approach is not being able to pass V_{DD}, while a PMOS only approach is not able to pass 0V. This is for similar reason as why PMOS makes a better "pull-up" device while NMOS makes a better "pull-down" device explained earlier. Without the ability to pass voltages rail-to-rail (to strongly turn on and off driven devices), slower transition (*i.e.* speed) of the driven gate is expected and static power dissipation potentially rises when devices are not completely off. This problem can be solved by using a transmission gate (having NMOS and PMOS together in parallel), as shown in Figure 5.10(c), but with increased device count. Relays, on the other hand, do not have such an issue. N-relay or P-relay only implementations are both effective to pass voltages rail-to-rail (Figures 5.10(d),(e)). Thus, large reduction in device count can be achieved using relays with the pass-gate architecture.



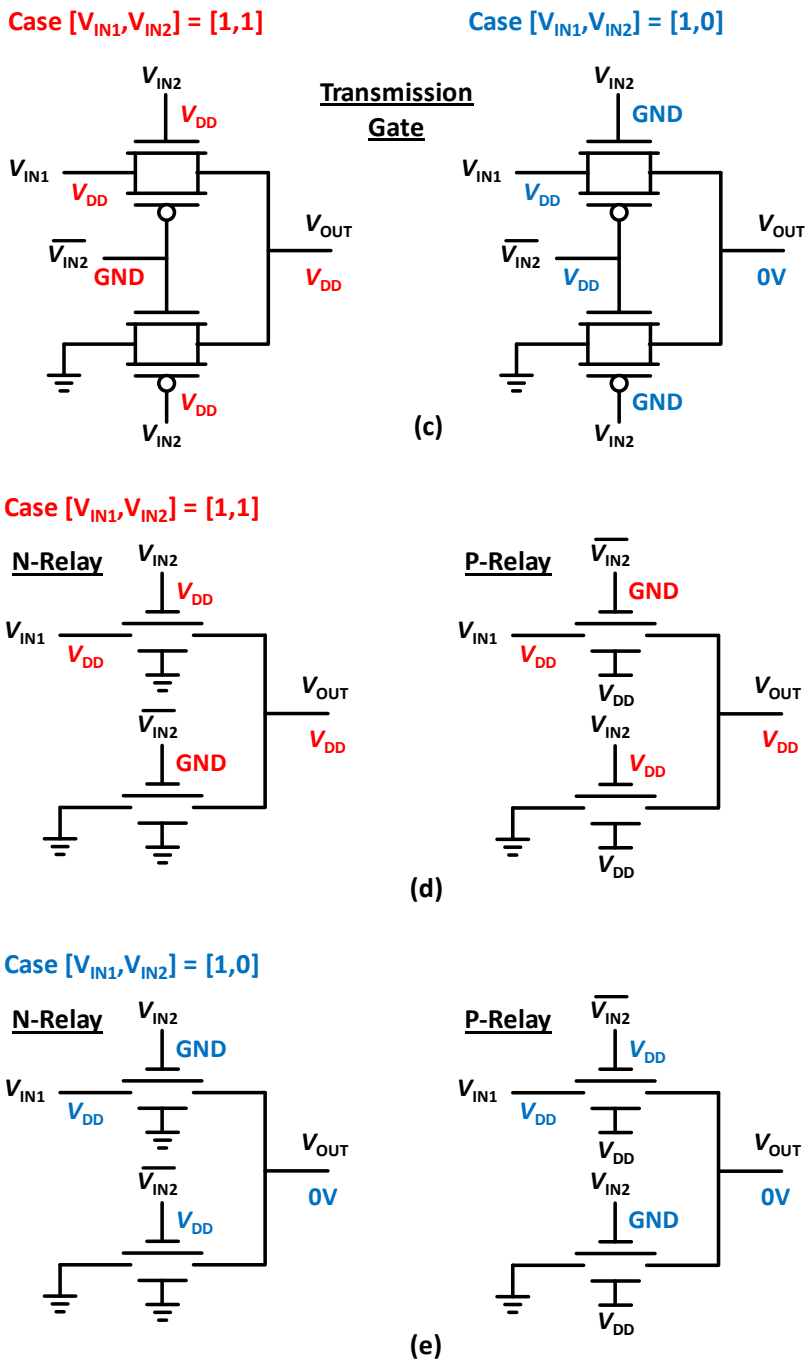


Figure 5.10: Comparison of pass-gate style AND gate implemented with CMOS and relays. Two cases, [V_{IN1},V_{IN2}]=[1,1] and [1,0] are compared. **(a)** Pure NMOS and pure PMOS implantation for [1,1] case and **(b)** [1,0] case. **(c)** Transmission gate implementation for [1,1] and [1,0] cases. **(d)** N-relay and P-relay implementation for [1,1] case and **(e)** [1,0] case.

5.4 Multi-Input/Multi-Output Relay Circuits

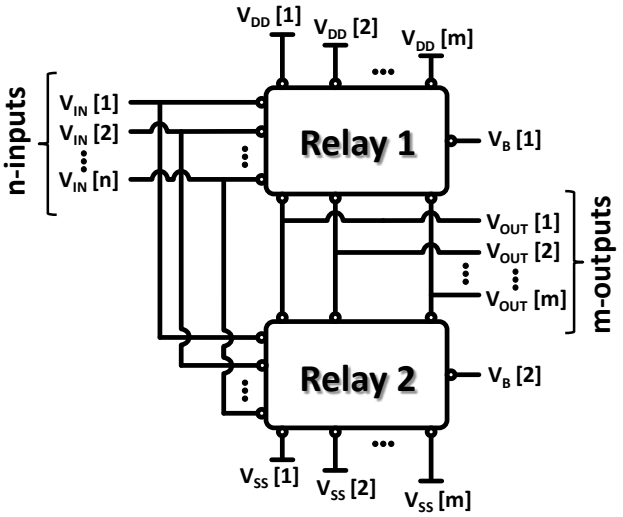


Figure 5.11: A generic multi-input, multi-output relay combinational logic circuit.

The ultimate, simplest (in terms of device count) implementation of any complex digital logic gate comprises only two switches: one "pull-up" switch that connects the output to the power supply when it is turned on, and one "pull-down" switch that connects the output to ground when it is turned on. The multi-input relay design (Section 4.5) allows a relay device to accommodate multiple inputs. In principle, almost any logic function can be implemented by adjusting the body bias levels and carefully designing the gate electrode areas to adjust the amount of electrostatic force each gate contributes to actuation. A generic complementary-relay logic circuit is illustrated in Figure 5.11. Each input signal is connected to one input electrode of the pull-up switch and also to one input electrode of the pull-down switch, and only one of these switches is on at any given time, *i.e.* they operate in a complementary manner. Note that each relay can comprise multiple pairs of source/drain electrodes as well, to provide for greater functionality, *e.g.* output signals at various voltage levels (as indicated in the figure) or differential output signals. To demonstrate these concepts for future zero-leakage digital ICs, the versatile functionality of circuits comprising only two multi-input, multi-output, electrostatically actuated relays is demonstrated in this work.

5.4.1 Single-Gate, Dual-Source Drain (1-Input, 2-Output) Relay Circuits

Figure 5.12 shows I_D-V_G curves of a single-gate, dual-source/drain relay with low hysteresis (<1 V). Note that the hysteresis voltage caused by surface adhesive force and the

pull-in phenomenon, and sets a lower limit for supply voltage scaling, even with body biasing. (To ensure that the relay is able to turn off, V_{RL} can be reduced to near 0 V at best, which means the lowest V_{PI} achievable is equal to the hysteresis voltage.) It is therefore crucial to lower hysteresis voltage for ultra-low power operation. With process and design optimization (Chapter 4), relays with <1 V hysteresis are achieved. Thus, V_{PI} can be reduced to <1 V with body biasing, for 1 V circuit operation in Figure 5.13.

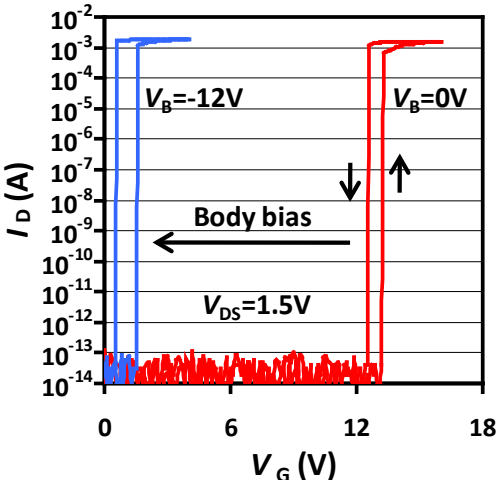


Figure 5.12: Measured I_D-V_G characteristics of a single-gate, dual-source/drain relay with <1 V hysteresis. With body biasing, V_{PI}, can be substantially reduced to <1 V with the relay still able to turn off reliably.

Two single-gate, dual-source/drain relays are connected together to form a dynamically configurable 1-input, 2-output complementary relay circuit, which can function either as an Inverter/Buffer or XOR/XNOR gate, depending on the electrode biasing configuration. Figure 5.13 shows Inverter/Buffer bias configuration. When the input voltage (V_{IN}) is high the top relay is off and the bottom relay is on, and vice versa. By connecting the sources to V_{DD} or GND, The left/right side source biases are V_{DD}/GND and GND/V_{DD} for the top and bottom relay respectively, such that complementary signals are achieved at the two sides. The relays used in this circuit have <1 V hysteresis, shown in Figure 5.12, so that low voltage (1 Volt) Inverter/Buffer operation is achieved with body biasing.

The same circuit is biased differently to achieve an XOR/XNOR gate (Figure 5.14). The XOR function makes use of the ambipolar nature of electrostatic force (actuation depends only on the magnitude of V_{GB}, not polarity). When the gate and body voltages are complementary, the top relay will turn on. When they are the same, the bottom relay will turn on. Since the body is used as an input electrode in this case, the input voltage range cannot be reduced by body biasing.

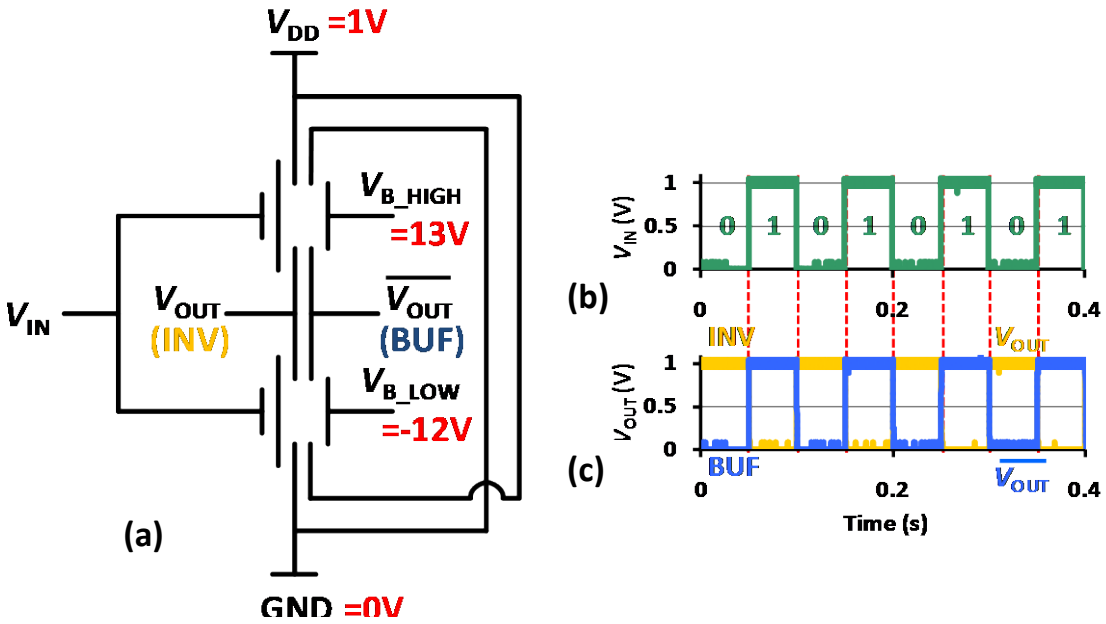


Figure 5.13: Dynamically configurable complementary relay logic circuit utilizing two single-gate, dual-source/drain relays. **(a)** Circuit schematic and bias configurations for Inverter/Buffer. V_{DD}=1V, V_{B_HIGH}=13V and V_{B_LOW}=-12V. **(b)** Measured input waveform. **(c)** Measured output waveforms.

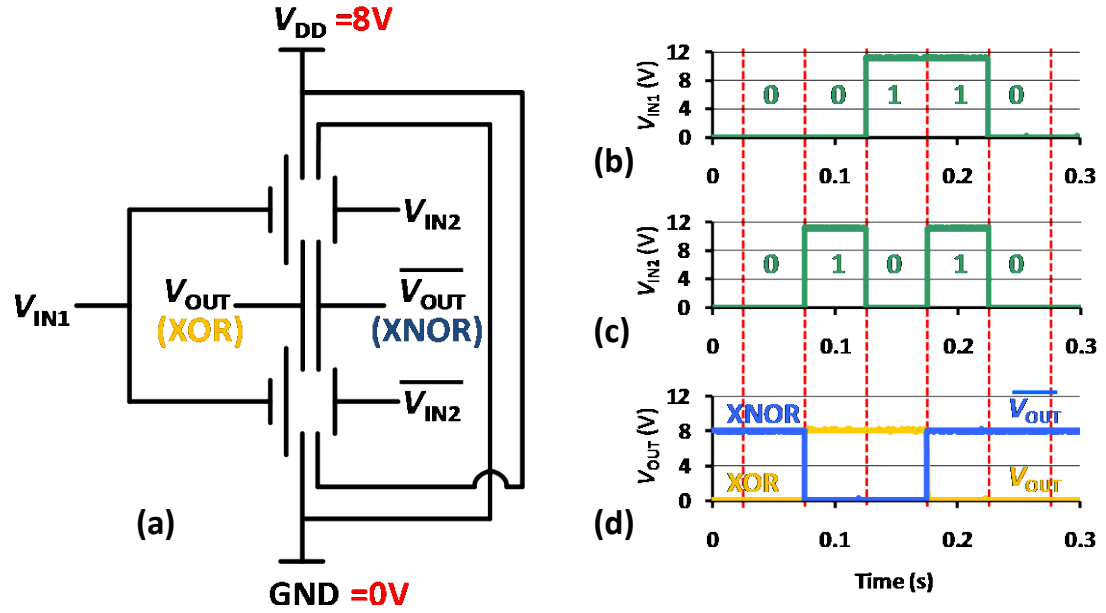


Figure 5.14: Dynamically configurable complementary relay logic circuit utilizing two single-gate, dual-source/drain relays. **(a)** Circuit schematic and bias configurations for XOR/XNOR. V_{DD} = 8 V. **(b)**, **(c)** Measured input waveforms. **(d)** Measured output waveforms.

5.4.2 Dual-Gate, Dual-Source/Drain (2-Input, 2-Output) Relay Circuits

Finally, two dual-gate, dual-source/drain relays are connected together to implement a dynamically configurable 2-input, 2-output complementary relay circuit, which can function either as an AND/NAND or OR/NOR gate, depending on the electrode bias configurations. Figure 5.15 shows AND/NAND functions. In this configuration, the top relay turns on only when all two gates are high. The bottom relay turns on when at least one gate is low. Again, switching is complementary and the source biases (V_{DD} or GND) determine whether that source will act as a "pull-up" or a "pull-down" connection. By simply adjusting the body bias voltages, the same circuit can implement OR/NOR functions, shown in Figure 5.16. The top relay now turns on when at least one gate is high. Similarly, the bottom relay now turns on only when all two gates are high.

This concept can be extended further to relay designs comprising greater than two input electrodes [6] and/or greater than two sets of source/drain electrodes, to implement more complex logic functions. Hence, multi-input/multi-input relays can implement multi-functional, dynamically configurable relay logic gates using only two relays, for highly compact implementation of relay-based integrated circuits.

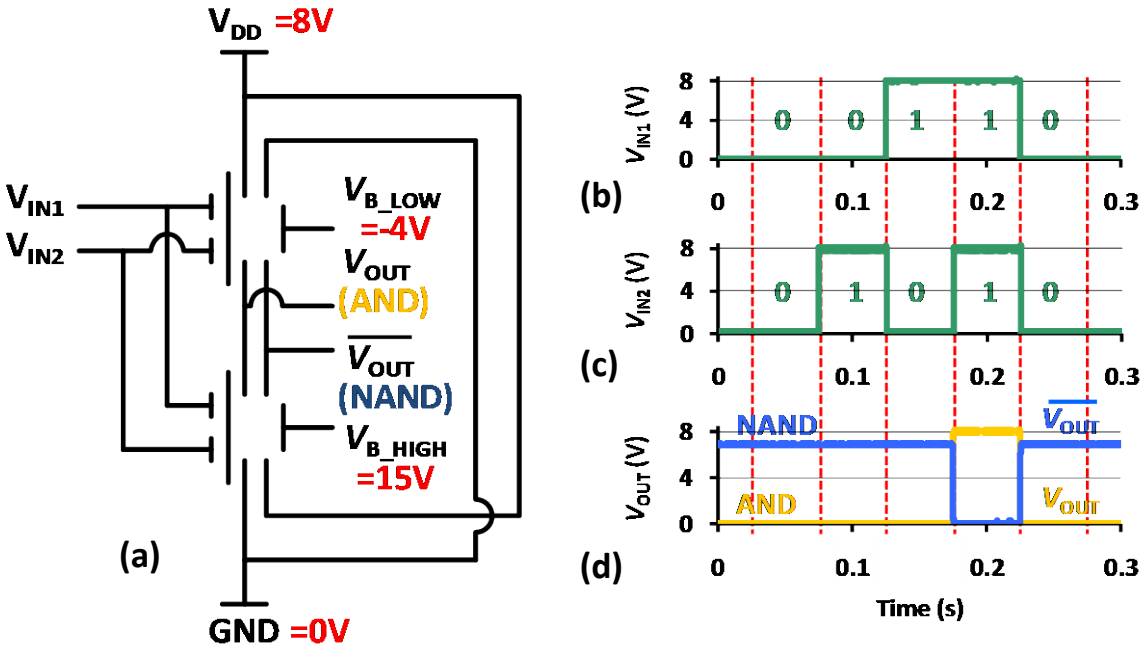


Figure 5.15: Dynamically configurable complementary relay logic circuit utilizing two dual-gate, dual-source/drain relays. **(a)** Circuit schematic and bias configurations for AND/NAND. V_{DD} = 8 V, V_{B_HIGH} = 15 V and V_{B_LOW} = -4 V. **(b)**, **(c)** Measured input waveform. **(d)** Measured output waveforms.

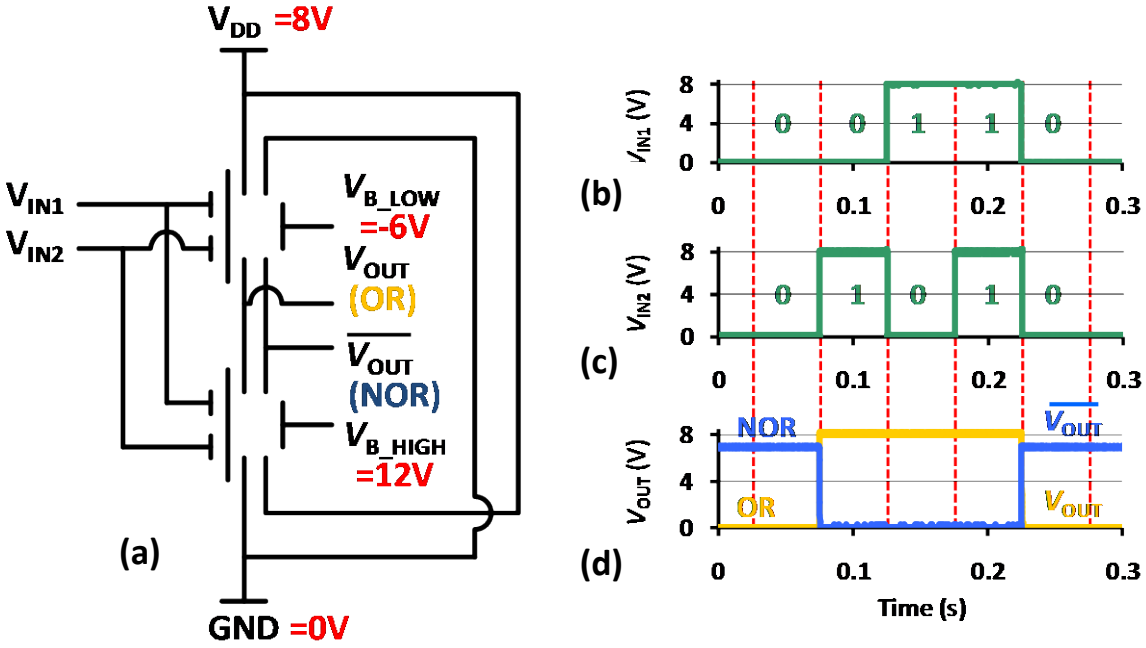


Figure 5.16: Dynamically configurable complementary relay logic circuit utilizing two dual-gate, dual-source/drain relays. **(a)** Circuit schematic and bias configurations for OR/NOR. V_{DD} = 8 V, V_{B_HIGH} = 12 V and V_{B_LOW} = -6 V. **(b)**, **(c)** Measured input waveform. **(d)** Measured output waveforms.

5.5 References

[1] K. Akarvardar, D. Elata, R. Parsa, G. C. Wan, K. Yoo, J. Provine, P. Peumans, R. T. Howe and H.-S. P. Wong, "Design considerations for complementary nanoelectromechanical logic gates," IEEE International Electron Devices Meeting Technical Digest, pp. 299-302, 2007.

[2] F. Chen, H. Kam, D. Markovic, T.-J. K. Liu, V. Stojanovic and E. Alon, "Integrated circuit design with NEM relays," IEEE/ACM International Conference on Computer-Aided Design, pp. 750-757, 2008.

[3] F. Chen, M. Spencer, R. Nathanael, C. Wang, H. Fariborzi, A. Gupta, H. Kam, V. Pott, J. Jeon, T.-J. K. Liu, D. Markovic, V. Stojanovic, and E. Alon, "Demonstration of integrated micro-electro-mechanical (MEM) switch circuits for VLSI applications," 2010 International Solid State Circuits Conference (San Francisco, California, USA), pp. 150-151, 2010.

[4] R. Nathanael, V. Pott, H. Kam, J. Jeon and T.-J. K. Liu, “4-terminal relay technology for complementary logic,” IEEE International Electron Devices Meeting Technical Digest, pp. 223-226, 2009.
[5] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolic, Digital integrated circuits, Prentice-Hall, 2003.
[6] J. Jeon, L. Hutin, R. Jevtic, N. Liu, Y. Chen, R. Nathanael, W. Kwon, M. Spencer, E. Alon, B. Nikolic, and T.-J. K. Liu, "Multi-input relay design for more compact implementation of digital logic circuits," IEEE Electron Device Letters, Vol. 33, No. 2, pp. 281-283, 2012.

Chapter 6
Conclusion

6.1 Summary

This dissertation begins with a perspective of CMOS technology scaling and the power crisis that has emerged for sub-100 nm technology nodes. Parallelism has been employed to alleviate this problem but is only a short-term solution since CMOS technology has a fundamental limit in energy efficiency that is related to the non-ideality of the transistor as an electronic switch. An alternative switch with more ideal behavior is therefore needed. The electrostatically actuated relay is proposed to meet this need. A relay has zero leakage current because the source and the drain electrodes are physically separated in the off state. In addition, since on/off switching is based on making and breaking physical contact, a relay's switching characteristic is hyper-abrupt (subthreshold swing (SS) is near zero), in contrast to a transistor's switching characteristic that is limited by the thermal voltage ($SS \geq 60$ mV/dec at room temperature).

Demonstration of micro-relay devices and circuits requires extensive process development. The 4-Terminal relay structure is found to be most attractive for digital integrated circuit (IC) applications, due to the benefits of having the body terminal. Apart from fixing the gate switching voltage with respect to the body voltage, it provides a means to electrically adjust the gate switching voltage, and as a result allows operation mimicking either an n-channel or p-channel MOSFET. A 4-Terminal relay structure consists of four materials that comprise the sacrificial, contact electrode, dielectric, and structural layers. Each of these needs to be properly optimized. For digital IC applications, high device reliability is of utmost importance, while slightly higher on-resistance is acceptable. Materials are selected based on the targeted device application, process integration requirements, and availability. Various process integration challenges are discussed and solutions proposed. The substrate temperature should not exceed 425°C for post-CMOS processing compatibility.

Based on the process development efforts, a robust 4-Terminal relay technology is developed and characterized to assess performance and reliability. Fabricated 4T relays exhibit good on-state current ($I_{ON} > 800$ μA for $V_{DS} = 1$ V) and zero off-state leakage current. Low-voltage switching (< 2 V) and low switching delay (100 ns) are demonstrated

by appropriately biasing the body terminal. Endurance exceeds 10^9 on/off cycles without stiction or wear issues. However, the 1st generation design suffers from parasitic electrostatic effects. Since capacitance forms between any two electrodes separated by a dielectric, the sizing of the electrodes needs to be optimized.

Next, an improved and more scalable process is presented. The design is also improved by optimizing the electrode sizing to eliminate parasitic effects. New features, such as extra pairs of source/drain, are incorporated into the design to enhance device functionality. The process flow is optimized by changing gap ratios and structural layer thickness. At best, V_{PI} ~10 V and ~12 V with the movable SiGe electrode and the fixed W electrode biased as gate, respectively. To complement process optimization, design optimization is presented by altering the dimensions, shape and orientation of the movable electrode, flexure and fixed electrode. Another 2 V reduction in V_{PI} can be expected without sacrificing reliability and yield. When fully optimized, the current 4T relay design and technology can operate reliably at ~8 V (SiGe gate) or ~10 V (W gate), with <1 V hysteresis.

Finally, simple relay-based circuits are demonstrated to assess the viability of this technology for digital ICs. A complementary inverter circuit is characterized and studied in depth to gain insight on the optimal body-biasing schemes and static noise margin for relay logic. Dynamically configurable logic gates are implemented using the 4-Terminal relay and the dual source/drain relay. All 2-input logic functions are demonstrated using only two relays with the multi-input/multi-output design. A low voltage (1 V) Inverter/Buffer circuit is achieved with body biasing. This work paves the way for compact implementation of highly compact relay logic circuits in the future.

6.2 Suggestions for Future Research

While relays are ideal switches with the potential to realize ultra-low-power integrated circuits, it has been difficult to scale down their operating voltage. The main challenge is strain gradient in the structural layer that causes out-of-plane deflection. Structures with strain gradient $<1\times10^{-4}$ μm^{-1} are required to address this challenge. Low-temperature poly-SiGe deposition has been studied extensively for MEMS applications requiring structural layer thicknesses >1 μm [1]. Since strain gradient is worse for thinner layers, optimization of the low-temperature poly-SiGe deposition process for structural layer thicknesses <1 μm is needed. Alternatively, multi-layered structural materials can be utilized to lower the overall strain gradient of the structure [2].

To demonstrate larger integrated circuits, such as a microcontroller, it is important for relays to operate reliably over $>10^{14}$ cycles. Tungsten is an excellent material to prevent contact failure from wear, plastic deformation, and welding-induced stiction. However, despite being intrinsically conductive, tungsten readily forms insulating native oxide (WO_3) in air. As a result, devices need to be "initialized" before testing by electrically breaking the contact native oxide. Native oxide also reforms readily during testing to make contact resistance unstable. This becomes a major challenge in implementing circuits, especially for large and complex ones, such as a microcontroller. To improve contact resistance stability, hermetic sealing technology [3] can be employed. An alternative material with more favorable contact properties may be needed. Ruthenium, for example, is an attractive candidate because it forms conductive oxide (RuO_2) in air, and still maintains relatively high hardness (albeit not as high as tungsten).

Another reason to optimize the contact material is to lower hysteresis voltage, for ultimately scaled relay operating voltage. Recall that the ultimate limit to voltage scaling in relays is the hysteresis voltage. Hysteresis is caused by pull-in mode operation and surface adhesive force. Devices can be made to operate in non-pull-in mode by reducing contact gap to 1/3 the actuation gap, in which case surface adhesion forces will be the sole contributor to hysteresis. Surface adhesive forces in TiO_2-coated W contacts have been studied and found to be worse than in pure W contact (WO_3 at the surface) [4]. A more extensive study needs to be done, especially as new materials are explored for contact material.

Finally, despite miniaturization, relays presented in this work are significantly larger than CMOS devices. For relays to be attractive as CMOS replacement device, a scaled technology will be required. Additionally, switching speed and operating voltage is expected to improve with technology scaling. With dimensional scaling comes many new processing challenges. As mentioned earlier, a new structural stack with low strain gradient need to be found. A new sacrificial layer deposition capability with more precise thickness control (such as ALD) is required to precisely define the gap thicknesses in the nanometer range. The ability to print small dimple sizes reliably is needed as well. If the current lithography capability is insufficient, e-beam lithography can be employed at the expense of throughput. Scaling the contact dimple size may give rise to new issues, such as a rising contact resistance. Relays with dimple size as small as 0.1μm×0.1μm have been fabricated and found to have higher contact resistance than those with 1μm×1μm dimple [5]. Effects of dimple size scaling on contact resistance and surface adhesion forces needs to be studied carefully. Relay technology scaling eventually will require more advanced manufacturing capabilities. There is much work yet to be done to realize relay-based ICs in mass production.

6.3 Outlook

Continued advancement in CMOS integrated circuit technology has become increasingly difficult due to a tradeoff between performance and energy efficiency. A power crisis has emerged due to off-state leakage that fundamentally limits CMOS energy efficiency. While new techniques, materials, and device architectures are being investigated to prolong CMOS scaling, the energy efficiency limit is governed by the physics of MOSFET operation, and exists regardless of materials or device architecture. Hence, an entirely new switch that consumes very low power when active and does not leak power when idle becomes necessary.

The benefits of having an extremely low power switch goes beyond high-performance computing. The recent shifting trend in the consumer market from personal computers (PCs) to mobile devices sparked a new era in computation that makes energy efficient computing devices all the more crucial. As we enter the era of ubiquitous computing, computing devices need to last a long time without the need to be frequently recharged, or to be self-sustaining by harvesting energy from the environment [6].

As the industry and researchers in academia start to look "beyond CMOS" for the next computing device, many have emerged. Some hold more promise than others. Tunnel field-effect transistors (TFET) that utilize the principle of band-to-band quantum mechanical tunneling for steep switching (<60 mV/dec) appears to be one of the more promising candidates [7]. More exotic devices employ a completely different paradigm without even having electrons as state variable. Nanoscale devices that pass tokens in spin, magnetic, photonic, excitonic, quantum, and heat domains have been proposed [8].

This dissertation proposes another promising approach in the search for a "beyond CMOS" switch alternative: nano-electro-mechanical relays for logic. Mechanical switches inherently have zero off-state leakage and extremely abrupt switching behavior. Hence, a relay is essentially the ideal switch, provided that its full potential can be achieved. While its speed and size is inferior to CMOS, reoptimized circuit design architectures can mitigate these penalties. Perhaps relays will not replace CMOS completely in areas where high speed/performance is most critical, but will be attractive for applications where energy efficiency is most important. A circuit-level assessment of energy-delay performance indicates that scaled relays can potentially provide for >10× improvement in energy efficiency as compared with CMOS, for applications requiring performance up to ~100 MHz clock frequency [9], [10]. In this dissertation, significant progress in relay technology development has been accomplished, but lowering the operating voltage and ensuring good reliability over the device lifetime remains a challenge. With further design and process improvements, a scaled nano-relay technology with optimized contact and structural materials, and potentially a proper hermetic sealing technology, can be compelling for energy-efficient information processing of the future.

6.4 References

[1] C. W. Low, “Novel processes for modular integration of silicon-germanium MEMS with CMOS electronics,” Ph.D. Dissertation, University of California, Berkeley, 2007.

[2] I. Chen, L. Hutin, C. Park, R. Lee, R. Nathanael, J. Yaung, J. Jeon and T.-J. K. Liu, "Scaled micro-relay structure with low strain gradient for reduced operating voltage," presented at the 221st ECS Meeting (Seattle, Washington, USA), May 2012.

[3] R. Candler, W. Park, H. Li, G. Yama, A. Partridge, M. Lutz, and T. Kenny, “Single wafer encapsulation of MEMS devices,” IEEE Trans. Adv. Packag., vol. 26, no. 3, pp. 227–232, Aug. 2003.

[4] D. Lee, V. Pott, H. Kam, R. Nathanael, T.-J. K. Liu, "AFM characterization of adhesion force in micro-relays," 2010 IEEE 23rd International Conference on Micro Electro Mechanical Systems, pp. 232-235, 2010.

[5] J. Yaung, University of California, Berkeley, unpublished work.

[6] M. Rahimi, H. Shah, G. S. Sukhatme, J. Heideman, and D. Estrin, “Studying the feasibility of energy harvesting in a mobile sensor network,” in Proc. IEEE International Conference on Robotics and Automation (ICRA'03), vol. 1, pp. 19-24, 2003.

[7] A. C. Seabaugh and Q. Zhang, “Low-voltage tunnel transistors for beyond CMOS logic,” Proceedings of the IEEE, vol. 98, no. 12, pp. 2095-2110, 2010.

[8] K. Bernstein, R. K. Cavin, W. Porod, A. Seabaugh, and J. Welser, “Device and architecture outlook for beyond CMOS switches,” Proceedings of the IEEE, vol. 98, no. 12, pp. 2169-2184, 2010.

[9] V. Pott, H. Kam, R. Nathanael, J. Jeon, E. Alon, and T.-J. K. Liu, "Mechanical computing redux: Relays for integrated circuit applications," Proceedings of the IEEE, vol. 98, no. 12, pp. 2076-2094, 2010.

[10] F. Chen, H. Kam, D. Markovic, T.-J. K. Liu, V. Stojanovic and E. Alon, “Integrated circuit design with NEM relays,” IEEE/ACM International Conference on Computer-Aided Design, pp. 750-757, 2008.